**BÁO CÁO KỸ THUẬT CHUYÊN SÂU: CÁC CHIẾN LƯỢC KIẾN TRÚC HIỆN ĐẠI ĐỂ ĐẠT HỘI TỤ NHANH VÀ ĐỘ CHÍNH XÁC CAO**

**1. Định Nghĩa Lại Hiệu suất Đào tạo: Chuyển đổi từ Epochs sang TTQ**

**1.1. Bối cảnh và Thách thức về Tốc độ Huấn luyện trong Kỷ nguyên Mô hình Lớn**

Thành công vượt trội của Mạng Nơ-ron Sâu (DNNs) trong các lĩnh vực như xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính dựa trên việc sử dụng các kiến trúc phức tạp và số lượng tham số khổng lồ.1 Tuy nhiên, sự phức tạp này kéo theo thách thức lớn nhất hiện nay: thời gian huấn luyện kéo dài và chi phí tính toán cao, đòi hỏi các thiết bị chuyên biệt như GPU và TPU.2 Do đó, động lực nghiên cứu cốt lõi đã chuyển dịch khỏi việc chỉ theo đuổi độ chính xác tuyệt đối sang tìm kiếm khả năng đạt được độ chính xác mục tiêu trong thời gian đồng hồ thực (wall-clock time) ngắn nhất.2

Nghiên cứu hiện đại coi các thuật toán có khả năng tăng tốc độ huấn luyện là một trong những mảng có giá trị nhất trong lĩnh vực học sâu.2 Điều này đòi hỏi sự tổng hợp giữa việc thiết kế kiến trúc thông minh và tối ưu hóa quy trình học tập, đảm bảo mô hình không chỉ học hiệu quả mà còn có khả năng mở rộng trên các hệ thống tính toán phân tán mà không làm suy giảm chất lượng gradient.

**1.2. Định lượng Hội tụ Nhanh: Time-to-Target Quality (TTQ)**

Trong môi trường nghiên cứu và công nghiệp, "hội tụ nhanh" không còn được đo bằng số lượng epoch mà bằng một chỉ số định lượng thời gian thực: Time-to-Target Quality (TTQ).

MLPerf Training benchmark suite là tiêu chuẩn công nghiệp để đo lường TTQ, xác định thời gian thực cần thiết để huấn luyện một mô hình đạt đến một mục tiêu chất lượng xác định.4 Tiêu chuẩn này liên tục được mở rộng để bao gồm các workloads tiên tiến, như LLM Fine-tuning (Llama3.1 8b), Graph Neural Networks (GNN, sử dụng mô hình R-GAT trên IGBH-Full), và Image Generation.4 Sự phát triển của MLPerf đã thúc đẩy những cải tiến đáng kể, với tốc độ huấn luyện trên chuẩn GPT3 Pretraining cho các hệ thống quy mô lớn (trên 10.000 bộ tăng tốc) được cải thiện hơn 3 lần trong vài năm qua.6

Sự hội tụ nhanh được đánh giá thông qua các metric hệ thống nâng cao. Ví dụ, Google đã giới thiệu khái niệm **Hiệu suất Scaling Hội tụ (Convergent Scaling Efficiency)**, một metric quan trọng ngoài thông lượng truyền thống, để đánh giá hiệu suất của TPU Trillium.7 Trillium đã chứng minh hiệu suất mỗi đô la (performance-per-dollar) tốt hơn 1.8 lần so với thế hệ TPU v5p trước đó, cùng với hiệu suất scaling ấn tượng 99%.7 Hiệu suất scaling hội tụ cho thấy, để giảm TTQ, mô hình không chỉ cần giảm số bước cần thiết mà còn phải duy trì hiệu quả cập nhật tham số khi mở rộng quy mô. Các thách thức trong huấn luyện song song phức tạp, chẳng hạn như việc sử dụng trọng số không nhất quán (inconsistent weights) trong các sơ đồ pipeline parallelism như "1F1B" (một thuật toán tối thiểu hóa bong bóng GPU) có thể ảnh hưởng đến hiệu quả cập nhật và làm giảm TTQ mặc dù tốc độ per-iteration có thể nhanh.8

**2. Đột Phá Kiến Trúc Cấp Độ Vĩ mô: Vượt qua Giới hạn Bậc Hai**

Kiến trúc hiện đại tập trung vào việc loại bỏ các điểm nghẽn vật lý và tính toán, đặc biệt là độ phức tạp bậc hai $O(N^2)$ theo chiều dài chuỗi $N$ và chiều mô hình $D$ vốn là đặc trưng của các khối Attention và MLP trong kiến trúc Transformer truyền thống.9

**2.1. Kiến trúc Sub-Quadratic và Monarch Mixer (M2)**

Monarch Mixer (M2) là một kiến trúc tiên phong giải quyết triệt để rào cản bậc hai này.9 M2 sử dụng **Ma trận Monarch (Monarch Matrices)**, một lớp ma trận có cấu trúc, vừa biểu cảm lại vừa hiệu quả phần cứng trên GPU.9

M2 đạt độ phức tạp dưới bậc hai, $O(p N^{(p+1)/p})$, trong đó $p$ là bậc của ma trận.9 Việc sử dụng cùng một primitive dưới bậc hai để trộn thông tin dọc theo cả chiều chuỗi và chiều mô hình mang lại những lợi ích đáng kể về TTQ và hiệu suất. Trong các thử nghiệm mô hình ngôn ngữ kiểu BERT (non-causal), M2 không chỉ khớp chất lượng trên GLUE mà còn sử dụng **ít hơn tới 27% tham số** so với các mô hình BERT-base/large.10 Quan trọng hơn, M2 đạt thông lượng cao hơn tới **9.1 lần** ở độ dài chuỗi $4K$ so với Transformer.10

Thành công của M2 cho thấy sự phức tạp bậc hai $O(N^2)$ không phải là một đặc điểm thiết yếu để đạt hiệu suất SOTA. Thay vào đó, việc thiết kế các ma trận có cấu trúc, hiệu quả tính toán có thể thay thế Attention và MLP, đồng thời giải phóng mô hình khỏi các giới hạn về độ dài ngữ cảnh và tăng đáng kể tốc độ đào tạo thực tế. Xu hướng này phù hợp với các hướng nghiên cứu khác nhằm giảm độ phức tạp Attention xuống tuyến tính $O(N)$ hoặc dưới bậc hai, bao gồm Linear Sequence Modeling và Sparse Sequence Modeling.11

**2.2. Tối ưu hóa Kiến trúc cho Thiết bị Biên**

Các kiến trúc tiết kiệm năng lượng, như Binary Spiking Neural Networks (BSNNs), rất quan trọng cho việc triển khai trên các thiết bị biên bị hạn chế tài nguyên.1 Tuy nhiên, những kiến trúc này đi kèm với các thách thức hội tụ riêng.

Việc huấn luyện BSNNs bị cản trở bởi trọng số nhị phân và hàm spike không khả vi, dẫn đến vấn đề **lật dấu trọng số thường xuyên (frequent weight sign flipping)**, làm chậm quá trình học.1 Để khắc phục điều này, các nhà nghiên cứu đã đề xuất **Cơ chế Điều chỉnh Gradient Thích ứng (Adaptive Gradient Modulation Mechanism - AGMM)**. AGMM hoạt động bằng cách điều chỉnh gradient một cách thích ứng để giảm tần suất lật dấu, cho phép BSNNs đạt được **tốc độ hội tụ nhanh hơn và độ chính xác cao hơn**, thu hẹp khoảng cách với các mô hình full-precision.1 Điều này chứng minh rằng, đối với các kiến trúc phi truyền thống và hiệu quả cao, cần phải thiết kế các thuật toán tối ưu hóa tùy chỉnh để đảm bảo TTQ tối ưu.

**2.3. Cải tiến Khối Xây dựng Cơ bản và Transformer Thị giác**

Nghiên cứu tiếp tục tối ưu hóa các khối xây dựng cơ bản. Công trình tại NeurIPS 2023 đã tái khẳng định Multi-Layer Perceptrons (MLPs), vốn thiếu thiên kiến cảm ứng thị giác (inductive bias), có thể đạt hiệu suất đáng kể trên các tác vụ thị giác (ví dụ: 58% trên ImageNet ReaL) bằng cách tăng quy mô.12 Do hiệu quả tính toán vốn có, các thí nghiệm pre-training quy mô lớn bằng MLP trở nên dễ tiếp cận hơn (có thể chạy trên GPU đơn).12

Trong lĩnh vực Vision Transformers (ViTs), kiến trúc hiện đại đã giải quyết điểm yếu của cơ chế self-attention truyền thống là thiếu bias không gian tường minh. Spatial Decay Transformer (SDT) giới thiệu cơ chế **Context-Aware Gating (CAG)**, tạo ra sự suy giảm không gian động, phụ thuộc vào nội dung, thay vì dựa trên các số liệu khoảng cách cố định.14 Bằng cách học cách điều chỉnh sự chú ý không gian dựa trên cả mức độ liên quan của nội dung và sự gần gũi về không gian, CAG giúp ViT cải thiện khả năng tổng quát hóa và đạt được sự hội tụ ổn định hơn trên các tác vụ phân loại và tạo sinh.14

**3. Kỹ thuật Ổn định và Hiệu quả Vi Mô: Đảm bảo Luồng Gradient Lành mạnh**

Các thành phần vi mô như lớp chuẩn hóa (normalization) và hàm kích hoạt (activation functions) là nền tảng để ổn định luồng gradient và do đó ảnh hưởng trực tiếp đến TTQ.

**3.1. Tối ưu hóa Chuẩn hóa: Từ BatchNorm đến RMSNorm**

Việc lựa chọn lớp chuẩn hóa đóng vai trò quan trọng trong việc tăng tốc độ hội tụ.

Ban đầu, Batch Normalization (BN), được giới thiệu vào năm 2015 15, được cho là khắc phục Internal Covariate Shift. Tuy nhiên, nghiên cứu gần đây cho thấy vai trò chính của BN là **làm mịn hàm mục tiêu (objective function smoothness)**, từ đó ổn định quá trình đào tạo và tăng tốc độ học.15 BN đặc biệt hiệu quả trong CNNs và khi sử dụng batch size lớn.16

Đối với các mô hình ngôn ngữ lớn (LLMs), Layer Normalization (LN) và một phiên bản đơn giản hóa, Root Mean Square Layer Normalization (RMSNorm), đã trở nên phổ biến. RMSNorm là một sự đơn giản hóa hiệu quả, chỉ thực hiện scaling dựa trên căn bậc hai trung bình (Root Mean Square) mà loại bỏ việc trừ giá trị trung bình (centering) và tham số bias $\beta$.17 Sự đơn giản này giúp **tăng cường ổn định huấn luyện** và **cải thiện hiệu suất tính toán** do giảm overhead, làm cho RMSNorm trở thành lựa chọn ưa thích (cùng với SwiGLU) trong các kiến trúc LLM hiện đại như LLaMA.17

Vị trí của lớp chuẩn hóa cũng rất quan trọng. Các LLM hiện đại áp dụng chiến lược **Pre-normalization**, trong đó chuẩn hóa được áp dụng *trước* các lớp self-attention và feed-forward network.17 Chiến lược này giúp giữ các activations trong phạm vi ổn định, ngăn chặn sự phát nổ hoặc biến mất của gradient, từ đó cải thiện tốc độ hội tụ tổng thể.17

**3.2. Lựa chọn Hàm Kích hoạt Tối ưu cho Tốc độ**

Hàm kích hoạt phải cân bằng giữa tốc độ tính toán per-iteration và khả năng duy trì luồng gradient ổn định để giảm tổng số epoch cần thiết.20

|  |  |  |  |
| --- | --- | --- | --- |
| **Hàm Kích hoạt** | **Tốc độ Per-Iteration** | **Số Epoch Cần thiết để Hội tụ** | **Đánh giá Hiệu quả TTQ** |
| ReLU | Nhanh nhất | Cao hơn (Nguy cơ Dying Units) | Có thể tăng Tổng thời gian Huấn luyện 21 |
| GELU | Trung bình (Nhanh hơn Swish ~15%) | Thấp hơn (Ổn định tốt) | Cân bằng tốt nhất giữa Tốc độ và Ổn định 21 |
| Swish | Chậm nhất (-20% đến -30% so với ReLU) | Thấp hơn (Gradient Flow tốt) | Thường là Tổng thời gian Huấn luyện dài nhất 21 |

Trong các mô hình Transformer như BERT và RoBERTa, ReLU mang lại tốc độ per-iteration nhanh nhất nhưng dễ gặp vấn đề "dying ReLU," có khả năng kéo dài tổng thời gian hội tụ. Ngược lại, GELU (Gaussian Error Linear Unit), là lựa chọn mặc định, cung cấp sự ổn định tốt hơn cho pre-training và thường yêu cầu ít epoch tổng thể hơn, tạo ra sự cân bằng tốt nhất giữa tốc độ và độ ổn định.21 Nghiên cứu tiếp tục khám phá các hàm kích hoạt mới như TeLU (Hyperbolic Tangent Exponential Linear Unit), nhằm thiết lập một tiêu chuẩn mới cho sự ổn định và hiệu quả học tập.22

**3.3. Tối ưu hóa Bộ nhớ và Suy luận Tốc độ Cao (GQA)**

Trong quá trình suy luận của Transformer, đặc biệt là khi xử lý ngữ cảnh dài, KV Cache (cặp Key/Value) tăng kích thước tuyến tính theo chiều dài ngữ cảnh, gây tắc nghẽn băng thông bộ nhớ.23 Sự tắc nghẽn này giới hạn batch size có thể phục vụ, làm giảm thông lượng chung và ảnh hưởng gián tiếp đến TTQ.

**Grouped Query Attention (GQA)** là một giải pháp cân bằng, nằm giữa Multi-Head Attention (MHA) và Multi-Query Attention (MQA).24 MHA tối đa hóa độ chính xác nhưng tăng chi phí băng thông bộ nhớ, trong khi MQA tối đa hóa tốc độ và hiệu quả bằng cách chia sẻ một bộ key/value vectors duy nhất qua tất cả các attention heads.23 GQA sử dụng một số lượng keys và values heads vừa phải ($G$, thường là $N/8$), giảm đáng kể bộ nhớ KV cache so với MHA.23 Việc giảm gánh nặng bộ nhớ này cho phép tăng batch size, từ đó tăng tốc độ suy luận decoder (MQA có thể nhanh hơn 12 lần và giảm cache 10-100 lần).24 GQA là một thành phần kiến trúc quan trọng, kết hợp với các kỹ thuật khác như Rotary Positional Embeddings (RoPE) và RMSNorm, để tối ưu hóa hiệu suất LLM.19

**4. Tối ưu hóa Động lực Học Nâng cao: Đẩy nhanh Quá trình Học**

Ngay cả với một kiến trúc tối ưu, tốc độ hội tụ cuối cùng vẫn phụ thuộc vào chiến lược tối ưu hóa (optimizer) và lịch trình tốc độ học (learning rate scheduling).

**4.1. Chính sách Tốc độ Học Chu kỳ và Hiện tượng Siêu Hội tụ**

Leslie N. Smith đã chứng minh rằng việc sử dụng **Tốc độ Học Chu kỳ (Cyclical Learning Rates)** và **Chính sách Một Chu kỳ (One-Cycle Policy - OCP)** có thể dẫn đến hiện tượng **Siêu Hội tụ (Super-Convergence)**, cho phép huấn luyện mạng nơ-ron nhanh hơn nhiều bậc so với phương pháp truyền thống.25

Cơ chế đằng sau Super-Convergence là việc sử dụng tốc độ học lớn hoạt động như một cơ chế **Điều chuẩn hóa (Regularization)** mạnh mẽ, giúp mô hình nhảy ra khỏi các điểm cực tiểu nông (sharp minima) và hướng tới các điểm cực tiểu phẳng (flat minima), có khả năng tổng quát hóa tốt hơn.25

OCP bao gồm một chu kỳ duy nhất, trong đó tốc độ học tăng tuyến tính từ giới hạn dưới (base\_lr) đến giới hạn trên (max\_lr), sau đó giảm trở lại.25 Giai đoạn cuối cùng là giai đoạn hủy diệt (annihilation phase), nơi LR giảm sâu hơn nữa, đảm bảo hội tụ vững chắc.25 Việc xác định các giới hạn LR tối ưu được thực hiện thông qua **LR Range Test**.25 Các schedulers hiện đại tiếp tục khai thác nguyên tắc này, ví dụ, Cyclical Log Annealing sử dụng "hồi phục độ dốc mạnh mẽ" (harsh restarting) để duy trì động lực học học tập mạnh mẽ.2

**4.2. Tối ưu hóa Thuật toán và Điều kiện hóa Gradient**

Các kỹ thuật tối ưu hóa đôi khi cần được tùy chỉnh cho các mục tiêu hội tụ cụ thể:

* **Tối ưu hóa Hai Pha:** Trong các nghiên cứu đòi hỏi độ chính xác cực cao (ultra-precision convergence), chiến lược tối ưu hóa hai pha, sử dụng Adam cho quá trình hội tụ nhanh ban đầu, tiếp theo là L-BFGS cho tinh chỉnh ổn định, cùng với cân bằng trọng số thích ứng, đã được chứng minh là hiệu quả.29
* **Điều kiện hóa Chuyên biệt:** Đối với các kỹ thuật nén mô hình như top-k sparsification, việc đạt được sự hội tụ ổn định đòi hỏi sự kết hợp của nhiều thủ thuật: kết hợp momentum vào phản hồi lỗi, gradient clipping, và giai đoạn warmup.30
* **Thách thức Scaling:** Khi mô hình đạt quy mô nghìn tỷ tham số, các thuật toán tối ưu hóa phải giải quyết các thách thức về scaling và hiệu quả bộ nhớ, bao gồm việc sử dụng adaptive momentum clipping.3

**4.3. Các Kỹ thuật Điều chuẩn hóa Dữ liệu**

Để đảm bảo tốc độ hội tụ nhanh dẫn đến khả năng tổng quát hóa cao (High Accuracy), các kỹ thuật điều chuẩn hóa dữ liệu vẫn là cần thiết. **Mixup** là một thủ tục điều chuẩn hóa mạnh mẽ, nội suy tuyến tính các cặp ví dụ huấn luyện và nhãn của chúng, làm giảm overfitting.31 Sự phát triển gần đây đã dẫn đến các biến thể chuyên biệt hóa theo kiến trúc, bao gồm PatchUp, GridMix, và **TokenMix** (được thiết kế cho Vision Transformers), nhằm tối ưu hóa sự tổng quát hóa cho các cấu trúc dữ liệu và mô hình mục tiêu cụ thể.32

**5. Khảo sát Định lượng và Các Học giả Tầm cỡ**

**5.1. Các Học giả Có ảnh hưởng Gần đây (2023-2025)**

Lĩnh vực tối ưu hóa và kiến trúc học sâu được dẫn dắt bởi các học giả có tầm ảnh hưởng lớn:

* **Các Nhà sáng lập:** Yoshua Bengio, người gần đây đã đạt 1 triệu trích dẫn trên Google Scholar 33, cùng với Yann LeCun và Geoffrey Hinton, tiếp tục là nền tảng cho lý thuyết học sâu.30
* **Nghiên cứu về Hiệu suất và Scaling:** Các nhóm nghiên cứu hiện đại tập trung vào việc vượt qua giới hạn của các kiến trúc hiện có. Leslie N. Smith là nhân tố cốt lõi đằng sau các động lực học học tập (Super-Convergence).26 Các nhà nghiên cứu như Gregor Bachmann, Sotiris Anagnostidis, và Thomas Hofmann đã cung cấp những hiểu biết sâu sắc về giới hạn của MLPs khi scaling và phát triển các chiến lược thích ứng để tăng tốc độ huấn luyện Vision Transformer (được công bố tại ICML 2024 và NeurIPS 2023).12
* **Hội nghị Nghiên cứu:** Các công bố tại các hội nghị hàng đầu như ICLR và NeurIPS năm 2024/2025 tiếp tục tập trung vào việc cải thiện tốc độ và đảm bảo hội tụ, bao gồm các nghiên cứu về "Flow matching achieves almost minimax optimal convergence" và "Improving Convergence Guarantees of Random Subspace Second-order Algorithm".36

**5.2. Vai trò của MLPerf trong Đo lường Hội tụ**

MLPerf Training không chỉ là một công cụ đo lường mà còn là một cơ chế định hướng, thúc đẩy ngành công nghiệp và học thuật tối ưu hóa các chiến lược hội tụ. Các kết quả của MLPerf cung cấp bằng chứng định lượng bằng TTQ về sự thành công của các tối ưu hóa kiến trúc và hệ thống. Ví dụ, kết quả MLPerf Training v4.0 đã minh họa hiệu suất vượt trội của các máy chủ sử dụng GPU NVIDIA H100 trong các workloads LLM và GNN, chứng minh khả năng của phần cứng hiện đại và các chiến lược tối ưu hóa đi kèm trong việc đạt được thời gian hội tụ mục tiêu.37 Các bài kiểm tra này liên tục phát triển, bao gồm cả các mô hình phức tạp như Llama 2 70B và DLRM-dcnv2.4

**6. Tổng kết Chiến lược Hội tụ Nhanh và Hướng Nghiên cứu Tương lai**

Hội tụ nhanh và độ chính xác cao trong học sâu hiện đại không phải là kết quả của một kỹ thuật đơn lẻ, mà là sự tổng hợp chiến lược (stack of optimizations) trên ba cấp độ chính: Kiến trúc Vĩ mô, Ổn định Vi mô, và Động lực Học Nâng cao.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cấp Độ Tối Ưu** | **Mục tiêu Chính** | **Chiến lược/Kỹ thuật Quan trọng** | **Tác động lên TTQ** |
| Kiến trúc Vĩ mô | Giảm độ phức tạp $O(N^2)$ | Monarch Mixer (M2), GQA, Linear Sequence Modeling | Tăng thông lượng per-iteration (9.1x cho M2), Giảm chi phí bộ nhớ KV Cache [10, 24] |
| Ổn định Vi mô | Đảm bảo luồng Gradient ổn định | RMSNorm + Pre-normalization, GELU, Batch Normalization (BN) | Tăng cường ổn định huấn luyện, làm mịn hàm mục tiêu, Giảm rủi ro gradient phát nổ [15, 17, 21] |
| Động lực Học | Giảm số bước cần thiết | One-Cycle Policy (Super-Convergence), AGMM (cho BSNNs), Tối ưu hóa hai pha | Giảm tổng số Epoch cần thiết để đạt mục tiêu [1, 25] |

Phân tích này cho thấy rằng:

1. **Thay thế Kiến trúc:** Việc tìm kiếm các kiến trúc dưới bậc hai (như M2) là trọng tâm để giải quyết triệt để vấn đề scaling và độ dài ngữ cảnh, cung cấp lợi thế về thông lượng per-iteration mà các tối ưu hóa phần mềm khó có thể đạt được.10
2. **Tối ưu hóa Thành phần Cốt lõi:** Việc áp dụng RMSNorm với Pre-normalization đã trở thành một tiêu chuẩn để đảm bảo sự ổn định của luồng gradient trong các mô hình Transformer sâu, một yếu tố then chốt để duy trì hiệu suất trong quá trình huấn luyện LLM quy mô lớn.17
3. **Khai thác Động lực Học Mạnh mẽ:** Việc sử dụng các chính sách tốc độ học tích cực và chu kỳ (OCP) là phương pháp đã được chứng minh để giảm đáng kể tổng thời gian hội tụ bằng cách giảm số lượng epoch cần thiết, đồng thời cung cấp tác dụng điều chuẩn hóa hiệu quả.25

**Hướng Nghiên cứu Tiếp theo**

Tương lai của hiệu suất huấn luyện sẽ tiếp tục tập trung vào việc vượt qua các giới hạn hệ thống và lý thuyết còn lại:

* **Phát triển Mô hình Phi-Attention và Sequence Modeling:** Tiếp tục đầu tư vào các kiến trúc dưới bậc hai (SSMs, M2) và các mô hình tuyến tính để giải quyết triệt để vấn đề ngữ cảnh dài, vốn là rào cản tính toán lớn nhất hiện nay.11
* **Tối ưu hóa Hệ thống và Tính nhất quán:** Cần các giải pháp kỹ thuật để cải thiện sự ổn định và hiệu suất trong các phương pháp đào tạo song song phức tạp, giải quyết các vấn đề như tính nhất quán của trọng số trong môi trường pipeline parallelism.8
* **Mở rộng Đảm bảo Hội tụ Lý thuyết:** Nghiên cứu cần mở rộng các đảm bảo lý thuyết về tốc độ hội tụ, hiện đang giới hạn trong chế độ khởi tạo, sang toàn bộ quỹ đạo huấn luyện để cung cấp hướng dẫn thiết kế kiến trúc mạnh mẽ hơn.38

**Nguồn trích dẫn**

1. Towards Accurate Binary Spiking Neural Networks: Learning with Adaptive Gradient Modulation Mechanism - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2502.14344v1>
2. cyclical log annealing as a learning rate scheduler - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/pdf/2403.14685>
3. Recent Advances in Optimization Methods for Machine Learning: A Systematic Review, truy cập vào tháng 10 30, 2025, <https://www.mdpi.com/2227-7390/13/13/2210>
4. Benchmark MLPerf Training | MLCommons Version 2.0 Results, truy cập vào tháng 10 30, 2025, <https://mlcommons.org/benchmarks/training/>
5. Introducing the MLPerf Training Benchmark for Graph Neural Networks - MLCommons, truy cập vào tháng 10 30, 2025, <https://mlcommons.org/2024/06/gnn-for-mlperf-training-v4/>
6. MLCommons MLPerf Training Expands with Llama 3.1 405B, truy cập vào tháng 10 30, 2025, <https://mlcommons.org/2025/05/training-llama31405b/>
7. Trillium MLPerf 4.1 training benchmarks | Google Cloud Blog, truy cập vào tháng 10 30, 2025, <https://cloud.google.com/blog/products/compute/trillium-mlperf-41-training-benchmarks>
8. Advances of Pipeline Model Parallelism for Deep Learning Training: An Overview - JCST, truy cập vào tháng 10 30, 2025, <https://jcst.ict.ac.cn/fileup/1000-9000/PDF/JCST-2024-3-4-3872-567.pdf>
9. MONARCH MIXER: A Simple Sub-Quadratic GEMM-Based Architecture - NSF PAR, truy cập vào tháng 10 30, 2025, <https://par.nsf.gov/servlets/purl/10494106>
10. Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture | OpenReview, truy cập vào tháng 10 30, 2025, [https://openreview.net/forum?id=cB0BImqSS9¬eId=98BZANkxc8](https://openreview.net/forum?id=cB0BImqSS9&noteId=98BZANkxc8)
11. Speed Always Wins: A Survey on Efficient Architectures for Large Language Models - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2508.09834v1>
12. Scaling MLPs: A Tale of Inductive Bias, truy cập vào tháng 10 30, 2025, <https://proceedings.neurips.cc/paper_files/paper/2023/file/bf2a5ce85aea9ff40d9bf8b2c2561cae-Paper-Conference.pdf>
13. Scaling MLPs: A Tale of Inductive Bias, truy cập vào tháng 10 30, 2025, <https://proceedings.neurips.cc/paper_files/paper/2023/hash/bf2a5ce85aea9ff40d9bf8b2c2561cae-Abstract-Conference.html>
14. Learning Spatial Decay for Vision Transformers - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2508.09525v1>
15. Batch normalization - Wikipedia, truy cập vào tháng 10 30, 2025, <https://en.wikipedia.org/wiki/Batch_normalization>
16. Batch vs Layer Normalization - Zilliz Learn, truy cập vào tháng 10 30, 2025, <https://zilliz.com/learn/layer-vs-batch-normalization-unlocking-efficiency-in-neural-networks>
17. RMSNorm in LLaMA: Enhance LLM Training Stability | Metric Coders, truy cập vào tháng 10 30, 2025, <https://www.metriccoders.com/post/rmsnorm-a-smarter-way-to-stabilize-your-llm-training>
18. Root Mean Square Layer Normalization - NIPS papers, truy cập vào tháng 10 30, 2025, <https://papers.nips.cc/paper/9403-root-mean-square-layer-normalization>
19. Flash Attention, RoPE And Other Performance Optimizers for LLMs - ResearchGate, truy cập vào tháng 10 30, 2025, <https://www.researchgate.net/publication/383532719_Flash_Attention_RoPE_And_Other_Performance_Optimizers_for_LLMs>
20. TeLU Activation Function for Fast and Stable Deep Learning - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2412.20269v1>
21. What is the impact of using ReLU, Swish, and Gelu activation functions on the training time of BERT and RoBERTa models? - Massed Compute, truy cập vào tháng 10 30, 2025, [https://massedcompute.com/faq-answers/?question=What%20is%20the%20impact%20of%20using%20ReLU,%20Swish,%20and%20Gelu%20activation%20functions%20on%20the%20training%20time%20of%20BERT%20and%20RoBERTa%20models?](https://massedcompute.com/faq-answers/?question=What+is+the+impact+of+using+ReLU,+Swish,+and+Gelu+activation+functions+on+the+training+time+of+BERT+and+RoBERTa+models?)
22. [2412.20269] TeLU Activation Function for Fast and Stable Deep Learning - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/abs/2412.20269>
23. Multi-head vs Multi-query vs Grouped-query attention | by Kantzuling - Medium, truy cập vào tháng 10 30, 2025, <https://medium.com/@kantzuling0307/multi-head-vs-multi-query-vs-grouped-query-attention-6981715eb6ec>
24. What is grouped query attention (GQA)? - IBM, truy cập vào tháng 10 30, 2025, <https://www.ibm.com/think/topics/grouped-query-attention>
25. How to achieve Super-Convergence and exploit One-Cycle policy: a simple guide - Medium, truy cập vào tháng 10 30, 2025, <https://medium.com/kirey-group/how-to-achieve-super-convergence-and-exploit-one-cycle-policy-a-simple-guide-430c1e0a3c1e>
26. [1506.01186] Cyclical Learning Rates for Training Neural Networks - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/abs/1506.01186>
27. Supercharge Training of your Deep Learning Models - Towards Data Science, truy cập vào tháng 10 30, 2025, <https://towardsdatascience.com/supercharge-training-of-your-deep-learning-models-7168ff81a042/>
28. davidtvs/pytorch-lr-finder: A learning rate range test implementation in PyTorch - GitHub, truy cập vào tháng 10 30, 2025, <https://github.com/davidtvs/pytorch-lr-finder>
29. [2507.20929] Breaking the Precision Ceiling in Physics-Informed Neural Networks: A Hybrid Fourier-Neural Architecture for Ultra-High Accuracy - arXiv, truy cập vào tháng 10 30, 2025, <https://arxiv.org/abs/2507.20929>
30. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks, truy cập vào tháng 10 30, 2025, <https://www.jmlr.org/papers/volume22/21-0366/21-0366.pdf>
31. MIXUP TRAINING AS THE COMPLEXITY REDUCTION - OpenReview, truy cập vào tháng 10 30, 2025, <https://openreview.net/pdf?id=xvWZQtxI7qq>
32. A curated list of awesome Mix - GitHub, truy cập vào tháng 10 30, 2025, <https://github.com/ChengtaiCao/Awesome-Mix>
33. Yoshua Bengio reaches 1 million citations on Google Scholar - UdeMNouvelles, truy cập vào tháng 10 30, 2025, <https://nouvelles.umontreal.ca/en/article/2025/10/24/yoshua-bengio-reaches-1-million-citations-on-google-scholar>
34. ICLR 2025 Papers, truy cập vào tháng 10 30, 2025, <https://iclr.cc/virtual/2025/papers.html>
35. Overview of MLPerf™ Training 4.0 Results - Dell Technologies Info Hub, truy cập vào tháng 10 30, 2025, <https://infohub.delltechnologies.com/en-au/p/overview-of-mlperf-tm-training-4-0-results/>
36. Quantitative convergence of trained single layer neural networks to Gaussian processes, truy cập vào tháng 10 30, 2025, <https://arxiv.org/html/2509.24544v2>