DeepSeek V3

Übersicht und wichtige Highlights

1. Modellname: DeepSeek-V3, ein Mixture-of-Experts (MoE) Sprachmodell mit 671 Milliarden Parametern, von denen 37 Milliarden pro Token aktiviert werden.
2. Trainingsdaten: Vorab trainiert auf 14,8 Billionen diversen, hochwertigen Tokens.
3. Kerninnovationen: Integriert Multi-Head Latent Attention (MLA) und DeepSeekMoE-Architekturen mit verlustfreier Hilfsverlustausgleich für Effizienz.
4. Trainingseffizienz: Erreicht vollständiges Training mit nur 2,788 Millionen H800-GPU-Stunden.
5. Kosteneffizienz: Die Trainingskosten werden auf 5,576 Millionen USD geschätzt, bei 2 USD pro GPU-Stunde.

Architekturinnovationen

6. Transformer-basierter Rahmen: Behält die Transformer-Architektur für Skalierbarkeit und Flexibilität.
7. Multi-Head Latent Attention (MLA): Reduziert die Inference-Speichernutzung durch Komprimierung von Schlüssel-Wert-Caches ohne Leistungsverlust.
8. DeepSeekMoE: Nutzt eine Kombination aus geteilten und gerouteten Experten für kosteneffizientes Training und hohe Recheneffizienz.
9. Hilfsverlustfreier Lastausgleich: Fügt Bias-Terms hinzu, um ausgeglichene Expertenlasten zu gewährleisten, ohne die Leistung zu beeinträchtigen.
10. Multi-Token-Prädiktion (MTP): Prädiziert sequenziell mehrere Tokens pro Position, verbessert die Dateneffizienz und die Vorabplanung der Darstellung.

Trainingsrahmen

11. FP8 Mixed Precision Training: Nutzt feingranulare Quantisierung und Speicherung in niedriger Präzision, um Speicher und Rechenleistung zu optimieren.
12. DualPipe-Algorithmus: Überlappt Rechen- und Kommunikationsphasen, reduziert Pipeline-Blasen und verbessert Parallelität.
13. Effiziente Cross-Node-Kommunikation: Nutzt optimierte Kernel für All-to-All-Operationen, nutzt NVLink- und InfiniBand-Bandbreiten.
14. Optimizer-Zustände in niedriger Präzision: Speichert Optimizer-Zustände in BF16, reduziert Speicherverbrauch ohne Leistungsverlust.
15. Speicheroptimierungstechniken: Berechnet bestimmte Operationen (z.B. RMSNorm) während der Rückwärtspropagation erneut, um Speicher zu sparen.

Vorab-Trainingsdetails

16. Stabiler Trainingsprozess: Keine irreparablen Verlustspitzen oder Rücksetzungen während des Vorab-Trainings.
17. Kontextlängenverlängerung: Kontextlänge auf 32K und anschließend auf 128K in zwei Stufen erweitert.
18. Trainingskosten: Vorab-Training benötigte 2,664 Millionen GPU-Stunden, Kontextverlängerung 119K GPU-Stunden und Nach-Training 5K GPU-Stunden.
19. Token-Effizienz: Trainingseffizienz durch Minimierung der GPU-Stunden pro Billion Tokens.
20. Hochwertige Daten: Vorab-Trainingsdaten für Vielfalt und Relevanz kuratiert.

Nach-Trainingsverbesserungen

21. Supervised Fine-Tuning (SFT): Richtet Modellausgaben an menschliche Präferenzen aus.
22. Verstärkungslernen (RL): Nutzt Group Relative Policy Optimization für Feinabstimmung.
23. Wissenstransfer: Integriert Schlussfolgerungsfähigkeiten von DeepSeek-R1-Modellen.
24. Steuerung des Ausgabe-Stils: Balanciert Genauigkeit mit Generationslänge und -stil.
25. Leistungsverfeinerung: Nach-Training verbessert Benchmark-Ergebnisse weiter.

Benchmark-Leistung

26. MMLU (Bildungsbenchmarks): Erreicht 88,5, übertrifft andere Open-Source-Modelle.
27. GPQA (Allgemeinwissen): Erreicht 59,1, vergleichbar mit GPT-4o und Claude-3.5-Sonnet.
28. Mathematische Benchmarks: Spitzenleistung in mathematischen Schlussfolgerungsaufgaben.
29. Code-Wettbewerbe: Exzelliert in Coding-Benchmarks wie LiveCodeBench.
30. Faktisches Wissen: Zeigt überlegene Ergebnisse in englischen und chinesischen Faktizitätsbenchmarks.

Inference und Deployment

31. Prefilling-Stufe: Kombiniert Tensor-Parallelismus (TP4), Sequenz-Parallelismus (SP) und Experten-Parallelismus (EP32) für Effizienz.
32. Decoding-Stufe: Nutzt EP320 mit IBGDA für niedrige Latenzkommunikation.

33. Dynamische Redundanz: Passt Expertenlasten dynamisch an, um Ressourcennutzung zu optimieren.
34. Trennung der Stufen: Prefilling- und Decoding-Stufen werden getrennt, um den Durchsatz zu erhöhen.
35. Hardware-Nutzung: Optimiert für H800-GPUs mit NVLink- und InfiniBand-Verbindungen.

Innovationen im Lastausgleich und Decoding

36. Bias-basiertes Routing: Fügt Bias-Terms hinzu, um dynamisch ausgeglichene Expertenlasten zu gewährleisten.
37. Spekulatives Decoding: Verbessert die Generationslatenz mit MTP-Modulen.
38. Redundante Experten: Dupliziert Experten mit hoher Last, um GPU-Arbeitslasten auszugleichen.
39. Node-beschränktes Routing: Beschränkt Token-Routing auf maximal 4 Knoten, um Kommunikationsaufwand zu reduzieren.
40. Kein Token-Dropping: Stellt sicher, dass alle Tokens während Training und Inference beibehalten werden.

Technische Details

41. Cluster-Konfiguration: Trainiert auf einem Cluster mit 2048 NVIDIA H800 GPUs.
42. Pipeline-Parallelismus: Nutzt ein 16-faches Parallelismus-Schema für Skalierbarkeit.
43. Speicherabdruck: Vermeidet kostspieligen Tensor-Parallelismus durch Optimierung der Speichernutzung.
44. Benutzerdefinierte Kernel: Entwickelt spezialisierte Kommunikationskernel, um Cross-Node-Operationen effizient zu handhaben.
45. Mixed Precision Optimierung: Kombiniert FP8- und BF16-Formate für optimale Trainingsdynamik.

Bewertung und Ergebnisse

46. Umfassende Benchmarks: Bewertet in diversen Domänen einschließlich Bildung, Coding und Schlussfolgerung.
47. Open-Source-Führerschaft: Geht als stärkstes Open-Source-Basismodell in seiner Kategorie hervor.
48. Vergleich mit Closed-Source-Modellen: Leistung vergleichbar mit GPT-4o und Claude-3.5-Sonnet.
49. Stärke im chinesischen Wissen: Übertrifft führende Modelle in chinesischen Faktizitätsbenchmarks.
50. Langkontext-Verarbeitung: Exzelliert in Aufgaben, die eine erweiterte Kontextverarbeitung erfordern.

Zukunftsrichtungen

51. Dynamische Redundanz-Erkundung: Untersucht adaptivere Redundanzstrategien.
52. Erweiterung des spekulativen Decodings: Erkundet weitere Anwendungen von MTP für Inference-Beschleunigung.
53. Hardware-Co-Design: Anpassung an nächste Generation von GPUs für verbesserte Leistung.
54. Breitere Benchmark-Abdeckung: Erweiterung der Bewertungen auf diversere Aufgaben.
55. Nachhaltigkeit: Weitere Reduzierung der Trainingskosten durch algorithmische und Hardware-Optimierungen.