DeepSeek V3

Vue d'ensemble et points forts clés

1. Nom du modèle : DeepSeek-V3, un modèle de langage Mixture-of-Experts (MoE) avec 671 milliards de paramètres, dont 37 milliards sont activés par jeton.
2. Jeu de données d'entraînement : Pré-entraîné sur 14,8 billions de jetons diversifiés et de haute qualité.
3. Innovations clés : Intègre l'attention latente multi-tête (MLA) et les architectures DeepSeekMoE avec un équilibrage de charge sans perte auxiliaire pour l'efficacité.
4. Efficacité de l'entraînement : Atteint un entraînement complet avec seulement 2,788 millions d'heures GPU H800.
5. Efficacité des coûts : Le coût de l'entraînement est estimé à 5,576M USD, en supposant 2 USD par heure GPU.

Innovations architecturales

6. Cadre basé sur Transformer : Conserve l'architecture Transformer pour la scalabilité et la flexibilité.
7. Attention latente multi-tête (MLA) : Réduit la mémoire d'inférence en compressant les caches clé-valeur sans perte de performance.
8. DeepSeekMoE : Utilise une combinaison d'experts partagés et routés pour un entraînement rentable et une efficacité computationnelle élevée.
9. Équilibrage de charge sans perte auxiliaire : Introduit des termes de biais pour maintenir des charges d'experts équilibrées sans compromettre la performance.
10. Prédiction multi-jeton (MTP) : Prédit séquentiellement plusieurs jetons par position, améliorant l'efficacité des données et la planification de la représentation.

Cadre d'entraînement

11. Entraînement en précision mixte FP8 : Utilise une quantification fine et un stockage en faible précision pour optimiser la mémoire et le calcul.
12. Algorithme DualPipe : Superpose les phases de calcul et de communication, réduisant les bulles de pipeline et améliorant le parallélisme.
13. Communication inter-nœuds efficace : Utilise des noyaux optimisés pour toutes les opérations de type all-to-all, utilisant les bandes passantes NVLink et InfiniBand.
14. États d'optimiseur en faible précision : Stocke les états d'optimiseur en BF16, réduisant la consommation de mémoire sans perte de performance.

15. Techniques d'optimisation de la mémoire : Recalcule certaines opérations (par exemple, RMSNorm) pendant la rétropropagation pour économiser de la mémoire.

Détails de l'entraînement

16. Processus d'entraînement stable : Aucune perte irréversible ou retour en arrière n'est survenu pendant le pré-entraînement.
17. Extension de la longueur du contexte : La longueur du contexte a été étendue à 32K puis à 128K en deux étapes.
18. Coûts d'entraînement : Le pré-entraînement a nécessité 2,664M heures GPU, l'extension du contexte 119K heures GPU et le post-entraînement 5K heures GPU.
19. Efficacité des jetons : L'efficacité de l'entraînement a été assurée en minimisant les heures GPU par billion de jetons.
20. Données de haute qualité : Le jeu de données de pré-entraînement a été soigneusement sélectionné pour sa diversité et sa pertinence.

Améliorations post-entraînement

21. Affinement supervisé (SFT) : Aligne les sorties du modèle avec les préférences humaines.
22. Apprentissage par renforcement (RL) : Utilise l'optimisation de la politique relative au groupe pour l'affinement.
23. Distillation des connaissances : Intègre les capacités de raisonnement des modèles DeepSeek-R1.
24. Contrôle du style de sortie : Équilibre la précision avec la longueur et le style de génération.
25. Affinement des performances : Le post-entraînement améliore encore les résultats des benchmarks.

Performance des benchmarks

26. MMLU (Benchmarks éducatifs) : Atteint 88,5, dépassant les autres modèles open-source.
27. GPQA (Connaissances générales) : Score de 59,1, comparable à GPT-4o et Claude-3.5-Sonnet.
28. Benchmarks mathématiques : Performance de pointe dans les tâches de raisonnement mathématique.
29. Concours de codage : Excellente performance dans les benchmarks de codage tels que LiveCodeBench.
30. Connaissances factuelles : Démontre des résultats supérieurs dans les benchmarks de factualité en anglais et en chinois.

Inférence et déploiement

31. Étape de pré-remplissage : Combine le parallélisme de tenseur (TP4), le parallélisme de séquence (SP) et le parallélisme d'experts (EP32) pour l'efficacité.
32. Étape de décodage : Utilise EP320 avec IBGDA pour une communication à faible latence.
33. Redondance dynamique : Ajuste dynamiquement les charges d'experts pour optimiser l'utilisation des ressources.
34. Séparation des étapes : Les étapes de pré-remplissage et de décodage sont séparées pour améliorer le débit.
35. Utilisation du matériel : Optimisé pour les GPU H800 avec interconnects NVLink et InfiniBand.

Innovations en équilibrage de charge et décodage

36. Routage basé sur le biais : Introduit des termes de biais pour assurer des charges d'experts équilibrées dynamiquement.
37. Décodage spéculatif : Améliore la latence de génération en utilisant des modules MTP.
38. Experts redondants : Duplique les experts à forte charge pour équilibrer les charges GPU.
39. Routage limité au nœud : Limite le routage des jetons à un maximum de 4 nœuds pour réduire la surcharge de communication.
40. Pas de suppression de jetons : Assure que tous les jetons sont conservés pendant l'entraînement et l'inférence.

Détails techniques

41. Configuration du cluster : Entraîné sur un cluster avec 2048 GPU NVIDIA H800.
42. Parallélisme de pipeline : Utilise un schéma de parallélisme 16 voies pour la scalabilité.
43. Empreinte mémoire : Évite le parallélisme de tenseur coûteux en optimisant l'utilisation de la mémoire.
44. Noyaux personnalisés : Développe des noyaux de communication spécialisés pour gérer efficacement les opérations inter-nœuds.
45. Optimisation de la précision mixte : Combine les formats FP8 et BF16 pour des dynamiques d'entraînement optimales.

Évaluation et résultats

46. Benchmarks complets : Évalué dans divers domaines incluant l'éducation, le codage et le raisonnement.

47. Leadership open-source : Émerge comme le modèle de base open-source le plus performant de sa catégorie.
48. Comparaison avec les modèles propriétaires : Performance comparable à GPT-4o et Claude-3.5-Sonnet.
49. Force dans les connaissances chinoises : Surpasse les modèles leaders dans les benchmarks de factualité en chinois.
50. Gestion des longs contextes : Excellente performance dans les tâches nécessitant un traitement de contexte étendu.

Directions futures

51. Exploration de la redondance dynamique : Enquête sur des stratégies de redondance plus adaptatives.
52. Expansion du décodage spéculatif : Explore d'autres utilisations de MTP pour l'accélération de l'inférence.
53. Co-conception matérielle : S'adapte aux GPU de prochaine génération pour une performance améliorée.
54. Couverture des benchmarks plus large : Étend les évaluations à des tâches plus diverses.
55. Durabilité : Réduit davantage les coûts d'entraînement grâce à des optimisations algorithmiques et matérielles.

Ce document fournit un résumé complet de DeepSeek-V3, encapsulant son architecture, ses méthodologies d'entraînement, ses performances de benchmarks et ses perspectives futures. Faites-moi savoir si vous avez besoin d'une élaboration sur des sections spécifiques ou de points supplémentaires !