Búsqueda Profunda V3

Descripción General y Puntos Destacados

1. Nombre del Modelo: DeepSeek-V3, un modelo de lenguaje Mixture-of-Experts (MoE) con 671 mil millones de parámetros, de los cuales 37 mil millones se activan por token.
2. Conjunto de Datos de Entrenamiento: Preentrenado con 14.8 billones de tokens diversos y de alta calidad.
3. Innovaciones Nucleares: Incorpora Multi-Head Latent Attention (MLA) y arquitecturas DeepSeekMoE con balanceo de carga sin pérdida auxiliar para eficiencia.
4. Eficiencia de Entrenamiento: Logra el entrenamiento completo con solo 2.788 millones de horas de GPU H800.
5. Eficiencia de Costos: El costo de entrenamiento se estima en 5.576M USD, asumiendo 2 USD por hora de GPU.

Innovaciones Arquitectónicas

6. Marco Basado en Transformer: Mantiene la arquitectura Transformer para escalabilidad y flexibilidad.
7. Multi-Head Latent Attention (MLA): Reduce la memoria de inferencia comprimiendo las caches de clave-valor sin pérdida de rendimiento.
8. DeepSeekMoE: Utiliza una combinación de expertos compartidos y enrutados para un entrenamiento rentable y alta eficiencia computacional.
9. Balanceo de Carga Sin Pérdida Auxiliar: Introduce términos de sesgo para mantener cargas de expertos equilibradas sin comprometer el rendimiento.
10. Predicción de Múltiples Tokens (MTP): Predice secuencialmente múltiples tokens por posición, mejorando la eficiencia de datos y la planificación de representación.

Marco de Entrenamiento

11. Entrenamiento de Precisión Mixta FP8: Utiliza cuantización de grano fino y almacenamiento de baja precisión para optimizar memoria y computación.
12. Algoritmo DualPipe: Superpone fases de computación y comunicación, reduciendo burbujas de pipeline e mejorando el paralelismo.
13. Comunicación Cruzada Eficiente entre Nodos: Emplea núcleos optimizados para operaciones de todos contra todos, utilizando ancho de banda NVLink e InfiniBand.
14. Estados del Optimizador de Baja Precisión: Almacena estados del optimizador en BF16, reduciendo el consumo de memoria sin pérdida de rendimiento.

15. Técnicas de Optimización de Memoria: Recomputa ciertas operaciones (por ejemplo, RMSNorm) durante la retropropagación para ahorrar memoria.

Detalles de Pre-Entrenamiento

16. Proceso de Entrenamiento Estable: No se produjeron picos de pérdida irrecuperables ni retrocesos durante el pre-entrenamiento.
17. Extensión de Longitud de Contexto: Extendió la longitud de contexto a 32K y posteriormente a 128K en dos etapas.
18. Costos de Entrenamiento: El pre-entrenamiento requirió 2.664M horas de GPU, la extensión de contexto 119K horas de GPU y el post-entrenamiento 5K horas de GPU.
19. Eficiencia de Tokens: La eficiencia de entrenamiento se aseguró minimizando las horas de GPU por billón de tokens.
20. Datos de Alta Calidad: El conjunto de datos de pre-entrenamiento fue curado para diversidad y relevancia.

Mejoras Post-Entrenamiento

21. Ajuste Fino Supervisado (SFT): Alinea las salidas del modelo con las preferencias humanas.
22. Aprendizaje por Reforzamiento (RL): Emplea Optimización de Políticas Relativas de Grupo para el ajuste fino.
23. Destilación de Conocimiento: Integra capacidades de razonamiento de los modelos DeepSeek-R1.
24. Control de Estilo de Salida: Equilibra precisión con longitud y estilo de generación.
25. Refinamiento de Rendimiento: El post-entrenamiento mejora aún más los resultados de las pruebas de referencia.

Rendimiento en Pruebas de Referencia

26. MMLU (Pruebas Educativas): Logra 88.5, superando a otros modelos de código abierto.
27. GPQA (Conocimiento General): Puntaje 59.1, comparable a GPT-4o y Claude-3.5-Sonnet.
28. Pruebas de Matemáticas: Rendimiento de vanguardia en tareas de razonamiento matemático.
29. Competencias de Código: Excelente en pruebas de codificación como LiveCodeBench.
30. Conocimiento Factual: Demuestra resultados superiores en pruebas de factualidad en inglés y chino.

Inferencia y Despliegue

31. Etapa de Prellenado: Combina paralelismo de tensores (TP4), paralelismo de secuencia (SP) y paralelismo de expertos (EP32) para eficiencia.
32. Etapa de Decodificación: Utiliza EP320 con IBGDA para comunicación de baja latencia.
33. Redundancia Dinámica: Ajusta dinámicamente las cargas de expertos para optimizar la utilización de recursos.
34. Separación de Etapas: Las etapas de prellenado y decodificación están separadas para mejorar el rendimiento.
35. Utilización de Hardware: Optimizado para GPUs H800 con interconexiones NVLink e InfiniBand.

Innovaciones en Balanceo de Carga y Decodificación

36. Enrutamiento Basado en Sesgo: Introduce términos de sesgo para asegurar cargas de expertos equilibradas dinámicamente.
37. Decodificación Especulativa: Mejora la latencia de generación utilizando módulos MTP.
38. Expertos Redundantes: Duplica expertos de alta carga para equilibrar las cargas de trabajo de la GPU.
39. Enrutamiento Limitado por Nodo: Restringe el enrutamiento de tokens a un máximo de 4 nodos para reducir la sobrecarga de comunicación.
40. Sin Eliminación de Tokens: Asegura que todos los tokens se retengan durante el entrenamiento e inferencia.

Detalles Técnicos

41. Configuración del Clúster: Entrenado en un clúster con 2048 GPUs NVIDIA H800.
42. Paralelismo de Pipeline: Emplea un esquema de paralelismo de 16 vías para escalabilidad.
43. Huella de Memoria: Evita el paralelismo de tensores costoso optimizando el uso de memoria.
44. Núcleos Personalizados: Desarrolla núcleos de comunicación especializados para manejar operaciones entre nodos de manera eficiente.
45. Optimización de Precisión Mixta: Combina formatos FP8 y BF16 para dinámicas de entrenamiento óptimas.

Evaluación y Resultados

46. Pruebas de Referencia Completas: Evaluado en diversos dominios incluyendo educación, codificación y razonamiento.

47. Liderazgo de Código Abierto: Emerge como el modelo base de código abierto más fuerte en su categoría.
48. Comparación con Modelos de Código Cerrado: Rendimiento comparable a GPT-4o y Claude-3.5-Sonnet.
49. Fuerza en Conocimiento en Chino: Supera a los modelos líderes en pruebas de factualidad en chino.
50. Manejo de Contexto Largo: Excelente en tareas que requieren procesamiento de contexto extendido.

Direcciones Futuras

51. Exploración de Redundancia Dinámica: Investigando estrategias de redundancia más adaptativas.
52. Expansión de Decodificación Especulativa: Explorando más usos de MTP para acelerar la inferencia.
53. Codiseño de Hardware: Adaptándose a las próximas generaciones de GPUs para un rendimiento mejorado.
54. Cobertura de Pruebas de Referencia Más Amplia: Expandiendo evaluaciones a más tareas diversas.
55. Sostenibilidad: Reduciendo aún más los costos de entrenamiento a través de optimizaciones algorítmicas y de hardware.

Este documento proporciona un resumen completo de DeepSeek-V3, encapsulando su arquitectura, metodologías de entrenamiento, rendimiento en pruebas de referencia y perspectivas futuras. ¡Háganme saber si necesita más detalles sobre secciones específicas o puntos adicionales!