I'm human

The DeepSeek V3 language model is a powerful Mixture-of-Experts (MoE) model that excels in various tasks. With 6,710 million parameters and each token activating 370 million parameters, it uses Multi-head Latent Attention (MLA) and deep MoE architecture to achieve efficient inference and cost-effective training. The model is trained on a massive dataset of 14.8 billion high-quality tokens and fine-tuned through supervised learning and reinforcement learning. Evaluations show that DeepSeek V3 outperforms other open-source models, rivaling the performance of leading closed-source models. The architecture of DeepSeek V3 consists of MLA, which enables efficient inference, and DeepSeekMoE, which facilitates cost-effective training. The model also employs a novel multi-token prediction (MTP) training target to enhance overall performance. A stability evaluation reveals that the training process is highly stable, with no irreversible loss peaks or rollbacks. The key components of the architecture are: 1. MLA: A framework that reduces memory usage and computation by compressing attention keys and values. 2. DeepSeekMoE: An architecture that uses expert gates to selectively apply different expert models to input tokens. The model's performance is showcased through various benchmark tests, including MATH-500, LiveCodeBench, and FRAMES. The Multiple Token Prediction (MTP) model, which predicts multiple tokens simultaneously, uses a Gloeckle-inspired approach to achieve this. The MTP module shares embedding and output layers with the main model, allowing for physical sharing of parameters and gradients. This shared architecture reduces memory usage and improves training efficiency. To further optimize memory usage, the team implemented several techniques, including: 1. Re-computing RMSNorm and MLA projections during backpropagation to avoid storing intermediate activations. 2. Using CPU-based Exponential Moving Average (EMA) to estimate model performance without increasing memory usage or computational overhead. 3. Deploying shallow layers of the MTP module on the same PP ordering as the main model, allowing for physical sharing of embedding and output layer parameters. The team also developed a fine-grained mixed precision framework for training DeepSeek-V3 using FP8 data format. This framework combines low-precision training techniques with strategic preservation of critical operations at original data format to balance training efficiency and numerical stability. In the mixed precision framework, most computation-intensive operations are performed at FP8 precision, while key operations are preserved at original data format. The framework is illustrated in Figure 6. The GEMM (General Matrix Multiply) operations are executed at FP8 precision, which accelerates model training theoretically by a factor of two. Additionally, the FP8 Wgrad GEMM allows for storing activations during backpropagation at FP8 format, reducing memory consumption significantly. Overall, the MTP module and mixed precision framework work together to optimize memory usage, reduce computational overhead, and improve the efficiency of DeepSeek-V3 training. **Section 3: Hardware and Software Implementations** The DeepSeek-V3 AI system uses an embedded module (BF16 or FP32) with an output head to process data. The system is designed to handle large-scale computing tasks while minimizing communication overhead. **Section 3.1: Embedded Module and Output Head** The embedded module is used for processing and the output head is used for sending results. The system uses a combination of techniques, including parallel processing, distributed memory, and optimized algorithms to improve performance. **Section 3.2: Communication Hardware** The communication hardware in DeepSeek-V3 is designed to overlap computation with communication, hiding latency and reducing dependencies on serial computing and communication bandwidth. However, this implementation relies on expensive Stream Multiprocessors (SM) which can limit compute throughput and lead to inefficiencies. **Recommendations for Hardware Design** 1. **Communication Hardware**: Develop hardware that offloads communication tasks from valuable compute units SM, making them available as GPU co-processors or network co-processors. 2. **Unified Interface**: Provide a unified interface for computing units to submit simple primitives-based communication requests across the entire IB-NVLink domain. **Section 4: Pre-training** The pre-trained model uses enhanced mathematical and programming samples with expanded language coverage (beyond English and Chinese). The data processing pipeline has been improved to minimize redundancy while maintaining dataset diversity. During the training process for DeepSeek-V3, we observed that the Fill-in-Middle (FIM) strategy can accurately predict middle text without harming its ability to predict the next token. This strategy was also used in the pre-training phase of DeepSeek-V2 and -V3. Specifically, we used a Prefix-Suffix-Middle (PSM) framework to structure the data, which involved inserting FIM markers into the input sequence. The PSM framework allowed us to incorporate FIM into the training process. We evaluated DeepSeek-V3's performance on various benchmark datasets, including MMLU, HellaSwag, PIQA, and more. We used two types of evaluation metrics: perplexity-based evaluation for certain datasets, and generative evaluation for others. For language modeling evaluation, we used Bits-Per-Byte (BPB) as a measurement standard. Our results show that DeepSeek-V3's baseline model outperformed its predecessor, DeepSeek-V2-Base, on most benchmark tests. We also compared our results to those of other state-of-the-art open-source baseline models, including Qwen2.5 72B Base and LLaMA-3.1 405B Base. Note that the evaluation framework has undergone changes in recent months, which may affect the performance of older models like DeepSeek-V2-Base. Overall, our findings demonstrate the effectiveness of FIM strategy in improving the performance of language models like DeepSeek-V3. The 405B Base model has become the strongest open-source model, outperforming others in various tasks. It was evaluated on multiple benchmarks, including MMLU, GPQA-Diamond, DROP, LongBench v2, and FRAMES. The results showed that DeepSeek-V3 is competitive with top-performing models like LLaMA-3.1-405B, GPT-4o, and Claude-Sonnet 3.5. It outperformed Qwen2.5 72B in MMLU-Pro and exceeded its competitors in MMLU-Redux. In the GPQA-Diamond benchmark, which simulates a PhD-level evaluation platform, DeepSeek-V3 achieved significant results, trailing only behind Claude 3.5 Sonnet. In the DROP benchmark, it scored an F1 score of 91.6 in the 3-shot setting, surpassing all other models. The text also describes the RL process used to optimize the model's performance. It mentions the use of group relative strategy optimization (GRPO) and the introduction of prompts from various domains, such as programming, mathematics, writing, role-playing, and question-answering. Overall, the results demonstrate DeepSeek-V3's strong ability to handle long-range context tasks and its competitiveness with top-performing models in multiple benchmarks. In 2024, DeepSeek-V3 outperformed its second-best model, Qwen2.5, by about 10 percentage points. The evaluation also included an open-ended generation task using LLM as the judge, with results shown in Table 7. Specifically, it followed the original configuration of AlpacaEval 2.0 (Dubois et al., 2024) and Arena-Hard (Li et al., 2024a), which used GPT-4-Turbo-1106 as the judge for paired comparisons. In Arena-Hard, DeepSeek-V3 achieved a win rate of over 86%. Similarly, it outperformed closed-source and open-source models in AlpacaEval 2.0, demonstrating its excellent ability in writing tasks and handling simple question-answer scenarios. Notably, it surpassed DeepSeek-V2.5-0905 by about 20 percentage points. In this section, the judging ability of DeepSeek-V3 is compared to that of GPT-4o and Claude-3.5. Table 8 shows the performance of these models on RewardBench (Lambert et al., 2024). DeepSeek-V3's performance was comparable to that of GPT-4o-0806 and Claude-3.5-Sonnet-1022, and it even surpassed other versions. Additionally, the judging ability of DeepSeek-V3 can be enhanced through voting technology. This section discusses the impact of distilled data on model contributions. The baseline model used short CoT (Chain of Thought) data, while the competitor used expert-checkpoint-generated data. Table 9 shows the effectiveness of distilled data, demonstrating significant performance improvements in LiveCodeBench and MATH-500 benchmark tests. An interesting trade-off was found: distillation improved performance but significantly increased average response length. To balance model accuracy and computational efficiency, careful parameter selection was performed for DeepSeek-V3's distillation. This section discusses self-reward mechanisms in reinforcement learning (RL). Self-reward plays a crucial role in guiding the optimization process. In some programming or mathematical scenarios, RL can exhibit extremely high efficiency. However, in more general scenarios, building feedback mechanisms becomes more challenging. Self-reward mechanisms become a possible solution, allowing models to reward themselves based on their generated content and guide the optimization process. This section evaluates the performance of DeepSeek-V3 using multi-token prediction (MTP) technology. Unlike traditional single-token prediction, MTP predicts the next two tokens. This method can significantly accelerate model decoding speed when combined with a predictive decoding framework (Leviathan et al., 2023; Xia et al., 2023). A natural question is whether the acceptance rate of additional predicted tokens is high. According to the results shown in Table 7, the acceptance rate is indeed very high. In conclusion, DeepSeek-V3 is a strong open-source language model that has outperformed its second-best model, Qwen2.5, by about 10 percentage points. Its training cost-effectiveness is high, thanks to FP8 training and fine-tuned engineering optimization. The distillation of reasoning abilities from the DeepSeek-R1 series models was also successful in the post-training phase. Overall assessment shows that DeepSeek-V3 has become one of the strongest open-source models and even surpasses some closed-source models in terms of performance, such as GPT-4o and Claude-3.5-Sonnet.