author - Visith Kumarapperuma

Deepseek V3: A Game-Changer in A.I. Here's Why It Matters

Currently, the AI models from the Chinese startup Deepseek are causing quite a stir in the AI space. Their latest reasoning model, Deepseek r1, shows better or equal performance to competitors. But above all, they achieved it with a fraction of the training and inference cost.

DeepSeek's AI Assistant overtook ChatGPT to become the most downloaded free app on the U.S. App Store. This development has led to market concerns about A.I. investments to major U.S. tech companies. Impacting share prices of tech firms including Nvidia.

So what made Deepseek such a big impact to A.I. ?

The significance of Deepseek as a disruptor in the industry lies in its approach. Unlike other companies that pushed for better hardware, Deepseek improved the algorithms. Thus achieving better results at a software level.

Note that the following details are for the Deepseek V3 model.

• Deepseek said it trained a model using a data centre of some 2,000 of Nvidia H800 GPUs.

• Time duration 2 months with the cost of the *final training run being ~$5.5 million

This ~$5.5M reflects the "rental" cost for the GPU hours needed to train DeepSeek-V3. It does not include:

1. The capital expenditure for owning the hardware.

2. Costs associated with prior research, ablation studies, or experiments on alternative architectures/algorithms/data.

Deepseek made training more efficient (45 times more efficient)

- Use 8-bit instead of 32-bit to save memory.

- Compress key value indices which eat up a lot of VRAM; they got 93% compression ratios.

- Do multi-token prediction instead of single-token prediction -> doubled inference speeds

- The MOE model decomposes a big model into small models that can run on consumer-grade hardware.

Summary of how Deepseek v3 was so efficient at training the frontier model

1. Model Architecture

The model employs a Mixture-of-Experts (MoE) architecture, where only 37B parameters fire for each token out of the total 671B. This sparse activation significantly reduces compute requirements compared to dense models.

The model uses Multi-head Latent Attention (MLA). This compresses the Key-Value cache, reducing memory usage and enabling more efficient training.

2. FP8 Mixed Precision Training:

They implemented an FP8 mixed precision training framework. Which reduces memory usage and accelerates training compared to higher precision formats.

Reduced memory footprint by up to 50% compared to traditional FP16/FP32 formats.

They use fine-grained quantisation strategies and increased accumulation precision to maintain accuracy.

3. Load Balancing Strategy

They pioneered an auxiliary loss-free strategy for load balancing in the MoE architecture. This improved performance without the drawbacks of traditional auxiliary loss methods.

4. Training Framework

They developed a custom training framework called HAI-LLM with several optimisations:

DualPipe algorithm for efficient pipeline parallelism. This reduces pipeline bubbles and overlapping computation and communication.

Efficient cross-node all-to-all communication kernels to fully utilise network bandwidth.

Careful memory optimisations to avoid using costly tensor parallelism.

Breakdown of the costs of the Deepseek v3 model

Deepseek's flagship model v3 showcases an architecture with a 671B parameter MOE (Mixture of Agents) with 37B active parameters per token

- Their success stems from breakthrough engineering: using MoE architecture, implementing FP8 mixed precision training, and developing a custom HAI-LLM framework.

- Deepseek excels at reasoning and math, surpassing GPT-4 and Claude 3.5 Sonnet.

- For writing and coding tasks, Claude 3.5 Sonnet maintains a slight lead.

- Deepseek pre-trained this model on 14.8 trillion high-quality data, taking 2,788,000 GPU hours on the Nvidia h800s cluster, costing around only $6 million

- the Llama 403b was trained on 11x of that, taking 30,840,000 GPU hours, also on 15 trillion tokens.

`So how true is the claim of $5.5 million, or is it another marketing trick?`

1. Underlying FLOP calculations

Model Details:

- Active Parameters: 37B (using FP8 precision)

- FLOPs per token: Using the rule of thumb “6 FLOPs per parameter per token.”

`37B×6 = 222B FLOPs per token`

- Total Training Tokens: Approximately 14.8 trillion tokens

- Total FLOPs required:

`222 B FLOPs/token×14.8 T tokens ≈ $3.3\times10^{24}$ FLOPs`

GPU FLOP Capacity (H800/H100):

An H100 is roughly estimated to deliver about.

3.958×10^{15} FLOPs (per second or per some standardised interval — here used as a comparative metric).

Ideal (Perfect Efficiency) GPU hours.

(Dividing total required FLOPs by per-GPU capability gives)

`$3.3\times10^{24}$ / $3.958\times10^{15}$ ≈ $8.33\times10^{8}$ seconds⇒≈0.4 million GPU hour`

Note: This “perfect efficiency” scenario is a lower bound. Real-world training is less efficient.

2. Adjusting for Real-World Inefficiencies (Comparison with Llama 3.1)

Reference Model: Llama 3.1 (405B parameters, 15 T tokens) reportedly required 30.84 M GPU hours in practice.

Recalculating FLOPs for Llama 3.1:

`Using the same math: $3.64\times10^{25}$ FLOPs required`

Scaling Efficiency

Using the ratio of FLOPs needed for DeepSeek-V3 versus Llama 3.1. and assuming similar inefficiencies.

The estimate adjusts to roughly 2.79M GPU hours for DeepSeek-V3 training.

3. DeepSeek-V3 Reported Training Breakdown

According to the DeepSeek-V3 paper

Pre-training Stage:

- Per Trillion Tokens: 180K H800 GPU hours

- Overall Pre-training: Total of 2,664K GPU hours

- This stage was completed in less than two months using a cluster of 2,048 H800 GPUs.

Context Length Extension:

- Additional 119K GPU hours

Post-training:

- An extra 5K GPU hours

Total GPU Hours:

`2,664 K+119 K+5 K≈2.788M GPU hours`

4. Cost Estimation

Assumed GPU Rental Price: $2 per GPU hour

Total Rental Cost:

`2.788M GPU hours×$2/hour≈$5.576 million`

as stated in Deepseek paper

During the pre-training stage, training DeepSeek-V3 on each trillion tokens requires only 180K H800 GPU hours… Consequently, our pre-training stage is completed in less than two months and costs 2664K GPU hours. Combined with 119K GPU hours for the context length extension and 5K GPU hours for post-training, DeepSeek-V3 costs only 2.788M GPU hours for its full training. Assuming the rental price of the H800 GPU is $2 per GPU hour, our total training costs amount to only $5.576M.

5. Summary

Theoretical (Perfect Efficiency) Estimate: ~0.4 M GPU hours (using idealised FLOP counts and assuming perfect hardware utilisation0

Adjusted (Real-World) Estimate (via Llama 3.1 comparison): ~2.79 GPU hours

DeepSeek-V3 Reported Breakdown:

Pre-training: 2,664K GPU hours

Context Extension: 119K GPU hours

Post-training: 5K GPU hours

Total: ~2.788 M GPU hours

Cost (at $2 per GPU hour): ~$5.576 million