Deep Seek

How is it different from other LLM systems?

Main source: https://en.wikipedia.org/wiki/DeepSeek

Engineering

The engineering of Deep Seek's various iterations has started from established LLM designs, and developed them to achieve similar performance with more basic hardware. Partly this is because of constraints on hardware purchasing, as the US restricts the sale of advanced Nvidia GPUs to China, but mostly it is the 'obvious' response of engineers to implementing a new generation of an established system – i.e. optimisation.

Deep Seek claims to use improved internal structures of its systems, with specialised decision layers that are activated on a case by case basis. It is not clear how much this relates to its LLM systems, and how much to its AI trading systems, which actually appear to be more central to the business. The Chat type systems are a spin-off of the trading systems, rather than the primary product.

Reduced precision arithmetic

One example of an optimisation is that the system apparently uses 8 bit floating point numbers instead of the more general 32 bit. This will save power, storage and improve speed.

Why this might work can be understood by considering the sigmoid function used in the nodes. Recall that the original neural network design used a true sigmoid to limit the range of the node output between -1 and 1. More recent LLM designs use simpler (to compute) functions which may be one-sided, such as computing the sum of inputs and then replacing any negative results with zero. The 8 bit floating point will presumably 'top out' and effectively limit large outputs, coming closer to the sigmoid behaviour.

Memory Architecture

Claims suggest that the memory architecture has been designed to be highly efficient, and avoid contention between different data layers, so that data movements are interlaced, and there are minimal delays waiting for data. I suspect that all the Chat type systems using dedicated hardware have spent a lot of effort on this, and Deep Seek probably has minor improvements in this area.

Challenges

The claimed reduced training costs are not accepted by some competitors, claiming that the full training costs have not been accounted for. Since the costs claimed by Deep Seek are about 16 times less, I think there are probably significant savings, but likely exaggerated.

Peter Whitham