Mixture of Experts & MoE in DeepSeek v2 and v3

February 27, 2025

MoE overview

- Mixture of Experts is an NN architectural design where data flows through subcomponents of the network in a variable fashion
- There are various forms of MoE that have been tried. We are going to talk about sparse MoE for transformer MLP layers
- Advantages of MoE include:
  - Greater expressive power for the same number of parameters
  - Less compute during inference than dense architectures
  - Faster training/convergence
- One disadvantage of MoE is needing more (GPU) memory during inference

Experts

- If you think about LLM layers performing tasks like:
  - Find the nouns
  - Is this word plural
  - Who designed the Eiffel Tower
  - Bunker Hill is not only a hill, but also a revolutionary war battle
- Then you will understand what different “experts” might do
  - WRONG – expert 1 knows biology, expert 2 knows philosophy, etc.
  - RIGHT – expert 1 processes many things, including digits and numbers,
    expert 2 processes many things, including a lot of verbs,
    expert 3 processes many things, include a lot of punctuation, etc.

MoE guide

- I found a wonderful resource online
- A Visual Guide to Mixture of Experts (MoE)
  By Maarten Grootendorst
  https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

Routing and resource utilization

- Because tokens in each layer are routed to a subset of experts, the number of parameters used to process each token is less than the total number of parameters
- If you have a server with many users, then different tokens for different inputs will, on average, use different experts
  - You will need many GPUs to store the weights for the different experts
  - By using *expert parallelism*, you can spread the work of many users across different GPUs and keep each one busy

DeepSeek use of MoE

- Starting with DeepSeek-v2, they have used MoE
- Kept a small number of experts that are always used
- Very large number of routed experts
- If using 8 routed experts, would like them to be located on as few devices as possible, so communication is reduced
  - DeepSeek-v2 encouraged desired distribution of experts by adding extra loss terms for: using all experts evenly, how many devices each token uses, and how much inter-device communication is required
  - DeepSeek-v3 switched from using the loss function to explicitly monitoring the load on each expert during training, and adding a bias to the routing to bump up/down how often each expert was used

MoE conclusion

- MoE architectures have been successfully implemented in LLMs, speeding up training and reducing compute cost during inference
- MoE increases the total amount of memory needed
  - Not helpful for running at home for a single user on a single GPU
  - Suited for high volume inferencing where many GPUs are available and the load from many users can be balanced across those GPUs
- DeepSeek has used MoE with a very high number of experts in both DeepSeek-v2 and DeepSeek-v3
- DeepSeek employed additional strategies to increase how often experts would be selected from the same device

References

- A Visual Guide to Mixture of Experts (MoE)
  Maarten Grootendorst
  https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts
- DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model
  DeepSeek-AI, et al. (2024)
  https://arxiv.org/abs/2405.04434
- DeepSeek-V3 Technical Report
  DeepSeek-AI, et al. (2024)
  https://arxiv.org/abs/2412.19437