DeepSeekMoE

Insop

DeepSeekMoE

DeepSeek LLM
Scaling Open-Source Language Models with Longtermism

Jan 2024

DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models

Damai Dai*[1,2], Chengqi Deng[1], Chenggang Zhao*[1,3], R.X. Xu[1], Huazuo Gao[1], Deli Chen[1], Jiashi Li[1], Wangding Zeng[1], Xingkai Yu*[1,4], Y. Wu[1], Zhenda Xie[1], Y.K. Li[1], Panpan Huang[1], Fuli Luo[1], Chong Ruan[1], Zhifang Sui[2], Wenfeng Liang[1]

DeepSeek-Coder: When the Large Language Model Meets Programming - The Rise of Code Intelligence

Daya Guo*[1], Qihao Zhu*[1,2], Dejian Yang[1], Zhenda Xie[1], Kai Dong[1], Wentao Zhang[1] Guanting Chen[1], Xiao Bi [1], Y. Wu[1], Y.K. Li[1], Fuli Luo[1], Yingfei Xiong[2], Wenfeng Liang[1]

April 2024

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2*†], Peiyi Wang[1,3*†], Qihao Zhu[1,3*†], Runxin Xu[1], Junxiao Song[1] Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

[1]DeepSeek-AI, [2]Tsinghua University, [3]Peking University

Jun 2024

DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model

DeepSeek-AI

research@deepseek.com

Dec 2024

DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

Jan 2025

deepseek

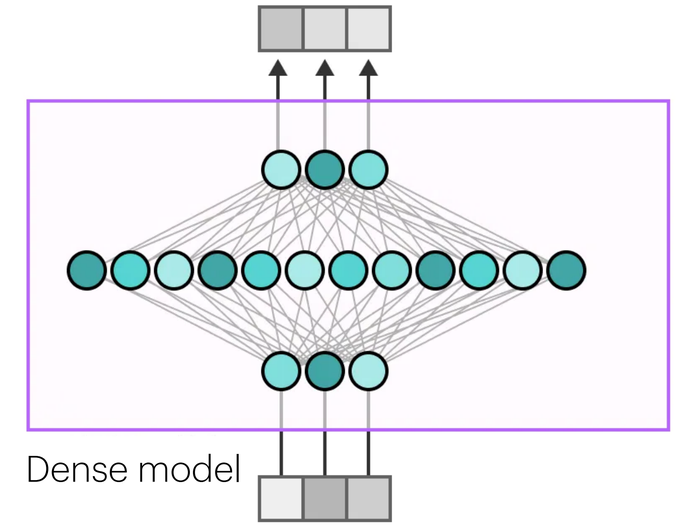
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

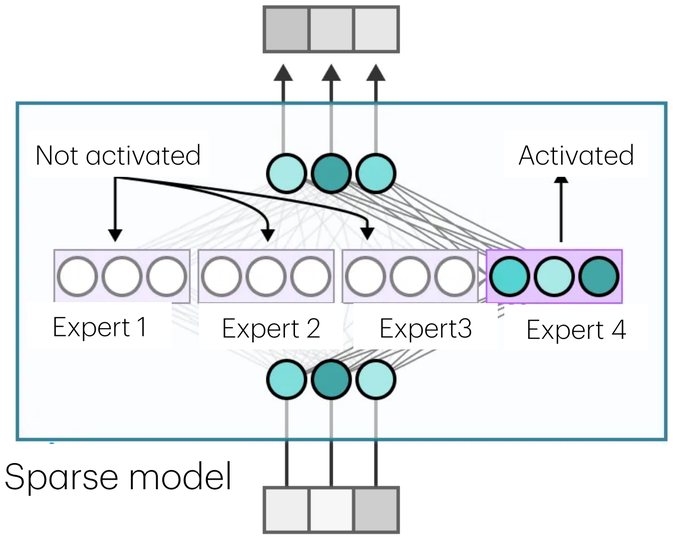
DeepSeek-AI

research@deepseek.com

Mixture of Expert, MoE (Review)

- According to scaling law*, model performance depends on i) # model parameters, N, ii) size of the dataset, D, iii) amount of compute in training, C
- MoE is a way to train larger model with lower training compute budget. Also MoE model requires lower inference budget
- MoE create a *sparse* model. Only 1 out of 4 experts is used. It is computationally efficient due to sparsity.
- MoE helps to increase parameters while keeping the compute cost manageable.





*: https://arxiv.org/abs/2001.08361 Scaling Laws for Neural Language Models

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

Mixture of Expert, MoE (Review)

- FFN layer of Transformer block is substituted with MoE layer
- Each expert is an identical structure FFN
- Challenges in MoE architectures
  - Knowledge hybridity: expert learns diverse knowledge due to its limited numbers
  - Knowledge redundancy: multiple experts learn common knowledge due to no common expert(s)
  - Routing collapse: selects a few experts so other experts may not trained fully

Output hidden

$$\mathbf{h}_t^l = \sum_{i=1}^{N} \left(g_{i,t}\, \mathrm{FFN}_i \left(\mathbf{u}_t^l \right) \right) + \mathbf{u}_t^l,$$

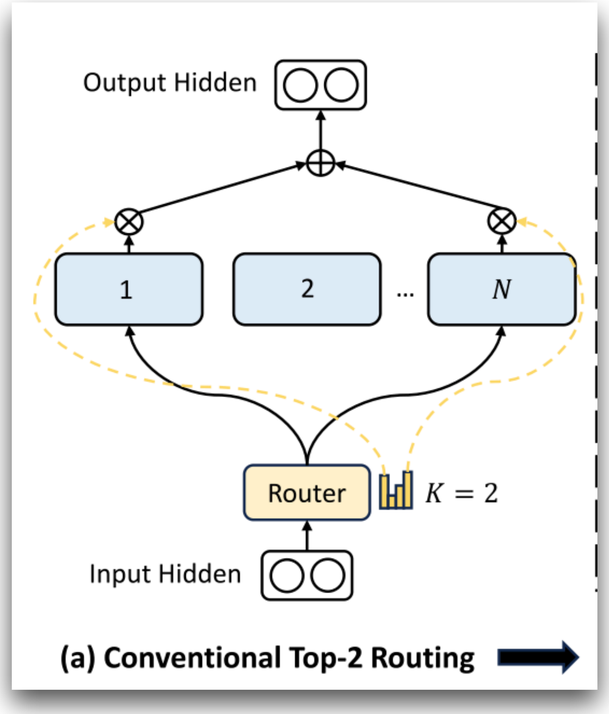
Gate, select expert

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N\}, K), \\ 0, & \text{otherwise}, \end{cases}$$

$$s_{i,t} = \mathrm{Softmax}_i \left(\mathbf{u}_t^{l\,T} \mathbf{e}_i^l \right),$$

Input hidden

Router

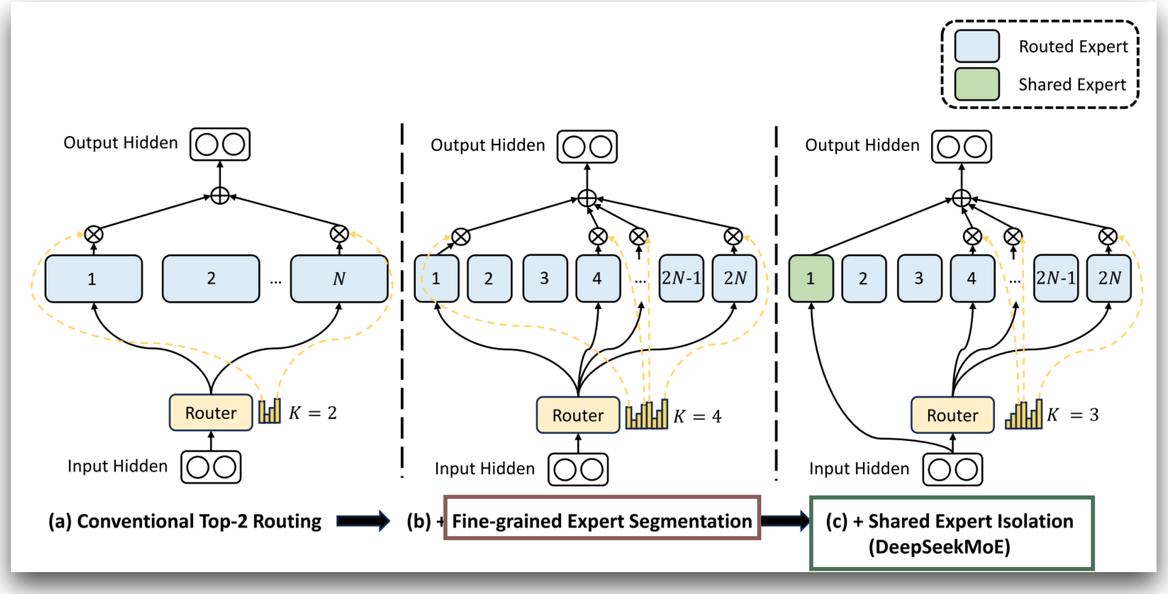


https://arxiv.org/abs/2401.06066 DeepSeekMoE

DeepSeekMoE

Design to improve expert specialization

- Fine-grained expert segmentation: splitting experts into finer groups and activating more experts allows them to learn diverse knowledge. Achieving more accurate and targeted knowledge acquisition.
- Shared expert isolation: Dedicate shared experts that are always activated to capture common knowledge across contexts. Prevent other experts from learning redundant knowledge.



Output hidden

$$\mathbf{h}_t^l = \sum_{i=1}^{K_s} \mathrm{FFN}_i\left(\mathbf{u}_t^l\right) + \sum_{i=K_s+1}^{mN} \left(g_{i,t}\,\mathrm{FFN}_i\left(\mathbf{u}_t^l\right)\right) + \mathbf{u}_t^l,$$

Gate, select expert

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | K_s + 1 \leqslant j \leqslant mN\}, mK - K_s), \\ 0, & \text{otherwise}, \end{cases}$$

Finer grained experts

Shared experts

$$s_{i,t} = \mathrm{Softmax}_i\left(\mathbf{u}_t^{l\,T} \mathbf{e}_i^l\right)$$

Input hidden

Router

https://arxiv.org/abs/2401.06066 DeepSeekMoE

DeepSeekMoE

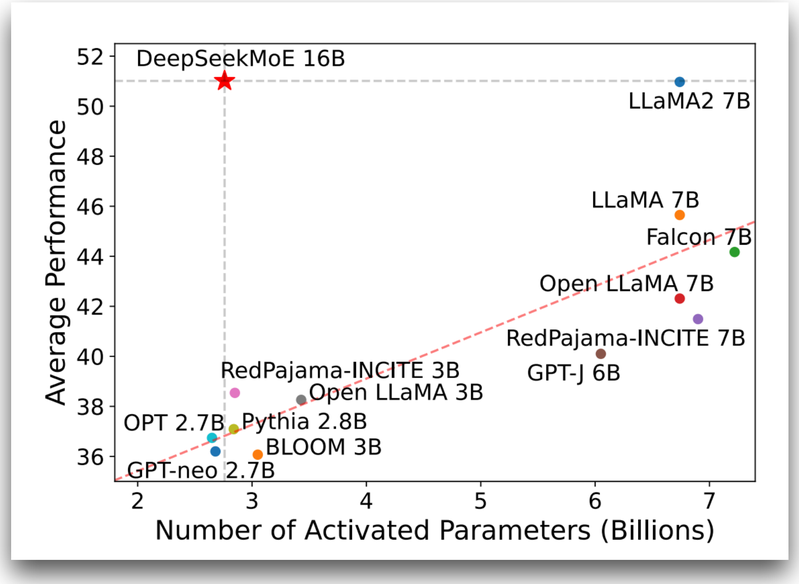
- To validate their idea, they used a small model with 2B parameters, DeepSeekMoE 2B
- DeepSeekMoE 2B outperform GShard 2.9B, which is large and x1.5 more activated experts.
- DeepSeekMoE 2B approaches close to DeepSeek 2B (dense model), which is a upper bound of MoE model

Metric	# Shot	Dense	Hash Layer	Switch	GShard	DeepSeekMoE
# Total Params	N/A	0.2B	2.0B	2.0B	2.0B	2.0B
# Activated Params	N/A	0.2B	0.2B	0.2B	0.3B	0.3B
FLOPs per 2K Tokens	N/A	2.9T	2.9T	2.9T	4.3T	4.3T
# Training Tokens	N/A	100B	100B	100B	100B	100B
Pile (Loss)	N/A	2.060	1.932	1.881	1.867	**1.808**
HellaSwag (Acc.)	0-shot	38.8	46.2	49.1	50.5	**54.8**
PIQA (Acc.)	0-shot	66.8	68.4	70.5	70.6	**72.3**
ARC-easy (Acc.)	0-shot	41.0	45.3	45.9	43.9	**49.4**
ARC-challenge (Acc.)	0-shot	26.0	28.2	30.2	31.6	**34.3**
RACE-middle (Acc.)	5-shot	38.8	38.8	43.6	42.1	**44.0**
RACE-high (Acc.)	5-shot	29.0	30.0	30.9	30.4	**31.7**
HumanEval (Pass@1)	0-shot	0.0	1.2	2.4	3.7	**4.9**
MBPP (Pass@1)	3-shot	0.2	0.6	0.4	0.2	**2.2**
TriviaQA (EM)	5-shot	4.9	6.5	8.9	10.2	**16.6**
NaturalQuestions (EM)	5-shot	1.4	1.4	2.5	3.2	**5.7**

Table 1 | Evaluation results for validation experiments. **Bold** font indicates the best. Compared with other MoE architectures, DeepSeekMoE exhibits a substantial performance advantage.

DeepSeekMoE

- DeepSeekMoE 16B achieves similar performance with DeepSeek 7B with only 40.5% computation!



Metric	# Shot	DeepSeek 7B (Dense)	DeepSeekMoE 16B
# Total Params	N/A	6.9B	16.4B
# Activated Params	N/A	6.9B	2.8B
FLOPs per 4K Tokens	N/A	183.5T	74.4T
# Training Tokens	N/A	2T	2T
Pile (BPB)	N/A	0.75	**0.74**
HellaSwag (Acc.)	0-shot	75.4	**77.1**
PIQA (Acc.)	0-shot	79.2	**80.2**
ARC-easy (Acc.)	0-shot	**67.9**	**68.1**
ARC-challenge (Acc.)	0-shot	48.1	**49.8**
RACE-middle (Acc.)	5-shot	**63.2**	61.9
RACE-high (Acc.)	5-shot	**46.5**	**46.4**
DROP (EM)	1-shot	**34.9**	32.9
GSM8K (EM)	8-shot	17.4	**18.8**
MATH (EM)	4-shot	3.3	**4.3**
HumanEval (Pass@1)	0-shot	26.2	**26.8**
MBPP (Pass@1)	3-shot	**39.0**	**39.2**
TriviaQA (EM)	5-shot	59.7	**64.8**
NaturalQuestions (EM)	5-shot	22.2	**25.5**
MMLU (Acc.)	5-shot	**48.2**	45.0
WinoGrande (Acc.)	0-shot	**70.5**	**70.2**
CLUEWSC (EM)	5-shot	**73.1**	72.1
CEval (Acc.)	5-shot	**45.0**	40.6
CMMLU (Acc.)	5-shot	**47.2**	42.5
CHID (Acc.)	0-shot	**89.3**	**89.4**