Model Editing

Thoughts on MoE Model Editing

密言

BUPT

2025 年 3 月 12 日

Abstract

Efficiently Editing Mixture-of-Experts Models with Compressed Experts

http://arxiv.org/abs/2503.00634

- 提出了“压缩专家”的概念，是作为完整专家的紧凑表示的轻量级模块
- 方法保留了最主要的专家，同时使用压缩专家替换其他辅助专家
- 压缩专家在各种任务中恢复了 90% 以上的性能，减少了 30% 的激活参数，同时节省了 20% 的推理成本

Intro

- 稀疏激活显著降低了与稠密模型相比的计算成本，同时保持了比较高的模型容量。
- MoE 模型的效率在很大程度上取决于激活专家的数量。
- 一些研究表明专家存在潜在的冗余性。

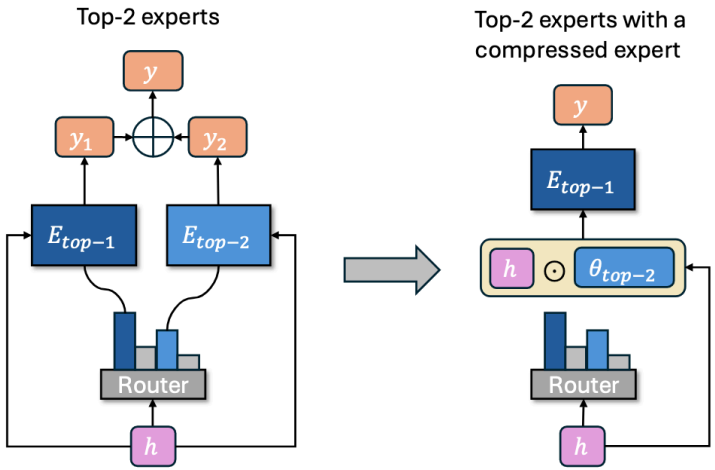


图 1: 左：MoE 架构，右：带压缩专家的 MoE 架构

Method

- 每次推理时，路由到 k 个专家，$n-k$ 个专家未被激活，其中 $k \ll n$。
- 对于每个 token，在激活的 k 个专家中，分为 k_m 个主要专家和 $k-k_m$ 个辅助专家。
- 辅助专家的参数量仅为完整专家的 0.05%。

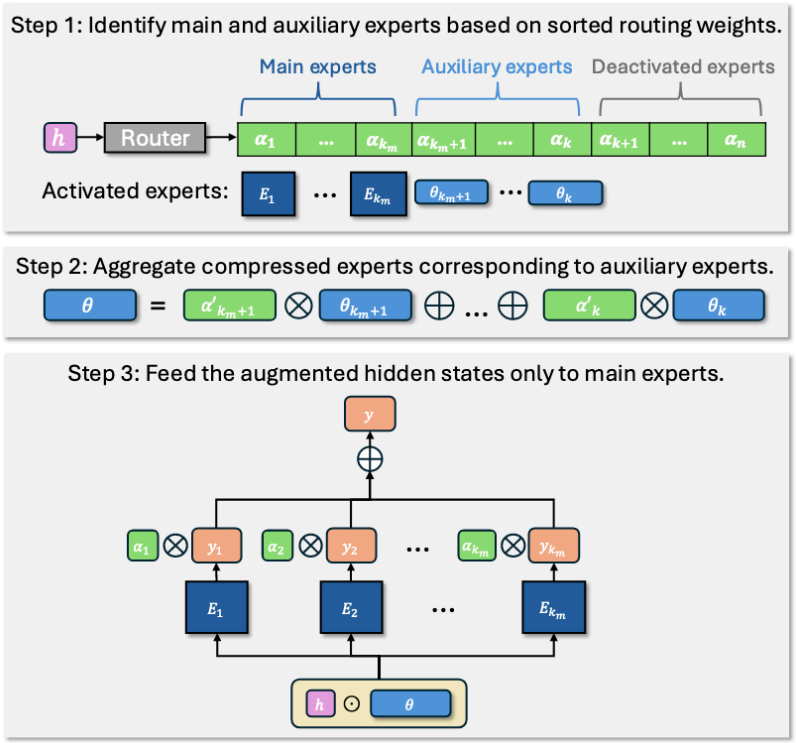


图 2: Step1：选择主要专家和辅助专家；Step2: 聚合辅助专家；Step3: 将增强的 hidden states 只输入主要专家

Experts Reduction 的探究

- 在超过一半的专家被设置为辅助专家时，模型性能的恢复能力会受到比较大的影响。
- 为了最大化效率同时不影响性能，使用平均分配的策略 $(k_m = k/2)$，保留 92% 的模型性能。

Configuration	w/o CE	w/ CE
Top-1	12.4	15.2
Top-2	18.0	20.6
Top-4	26.8	29.7
Top-8	32.3	/

Table 1: GSM8K 0-shot CoT exact match scores (%) for OLMoE with varying numbers of main experts. Compressed experts (CE) improve performance across reduced configurations, but cannot fully recover the performance lost when reducing more than half of the experts.

Experiments

1. 固定一定数量的专家进行 SFT
2. 在 SFT 过后，辅助专家的冗余性会变得更加明显，此时进行辅助专家的压缩
3. 实验结果表明，压缩专家后平均性能保留 94% 左右

Task (→) Metric (→)	IFEval 0-shot Loose Acc	BBH 3-shot EM	TruthfulQA MC2	GSM8K 0-shot CoT EM	HumanEval 0-shot Pass@1	Avg (↑)	Latency (↓) s
Pretrained	25.3	63.1	45.8	31.8	48.0	42.8	-
Top-2 SFT	54.9	69.5	49.3	76.7	67.1	63.5	5.59
Top-1 SFT	51.4	67.1	49.2	57.6	56.9	56.4	4.01
Top-1 SFT w/ CE	53.6	67.3	48.8	65.5	64.2	60.0	4.35
Norm. Perf. (%)	97.6	96.8	99.0	85.4	95.7	94.5	-

Table 3: Phi-MoE (pretrained with 2 activated experts) results on general tasks. The inference latency is measured by the time required to process a fixed number of randomly generated tokens in forward passes. Normalized performance measures the relative performance with respect to the full-expert configuration.

Task (→) Metric (→)	IFEval 0-shot Loose Acc	BBH 3-shot EM	TruthfulQA MC2	GSM8K 0-shot CoT EM	HumanEval 0-shot Pass@10	Avg (↑)	Latency (↓) s
Pretrained	16.5	32.1	35.8	12.1	18.7	23.0	-
Top-8 SFT	39.6	32.5	41.1	36.9	39.9	37.9	7.14
Top-4 SFT	34.2	30.7	39.2	33.1	36.6	34.8	5.31
Top-4 SFT w/ CE	35.1	31.5	41.4	35.9	38.4	36.5	5.83
Norm. Perf. (%)	88.6	96.9	100.7	97.3	96.2	96.3	-

Table 4: OLMoE (pretrained with 8 activated experts) results on general tasks.

MoE 架构

传统的 tranformer 模型中，将每个 FFN 层替换成 MoE 层，MoE 层由门控网络（Gate/Router）和专家网络（Experts）构成。MoE 模型在增大参数量的同时，由于稀疏激活，提高了训练和推理效率。

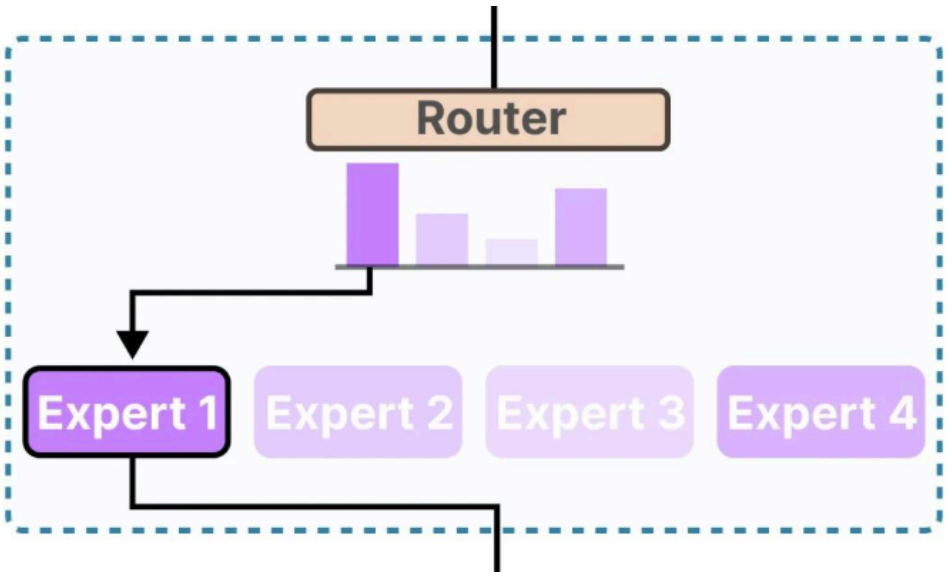


图 3: MoE 层的结构

专家的自发专化和动态路由策略

自发专化

- 在训练初期，门控网络参数随机，各专家接受数据均匀
- 后期的局部梯度更新时某些专家专注于处理特定类型的数据

动态路由策略

- 正反馈循环：将某个特定数据分配给表现优异的专家。
- 专家和路由网络的协同演化：专家获得专长，门控网络根据反馈更新参数，使路由更加精准。

Top-K 路由策略

- 噪声注入：在路由得分上面加入随机噪声，使得初期路由策略具有随机性，有利于均衡训练。
- 保留 Top-K：只选取得分最高的 k 个专家，其余专家不激活。

对专家“偏好”的探讨

一次交互有多个 token，每个 token 的路由决策一样吗？

每个 token 的路由决策都是独立的。

不同专家的偏好是怎样的？

不同领域和不同 token 都有一定的专家偏好。

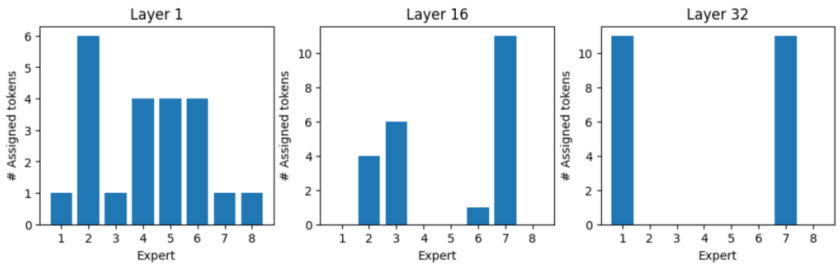


图 4: token "who" 的专家分配（结果来自 Mixtral 8*7B）

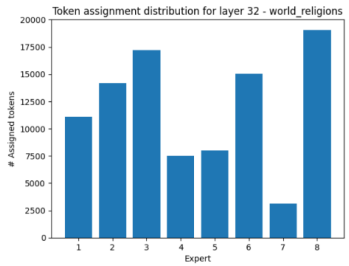


图 5: 世界宗教主题在 32 层的 expert 的 token 分配

使用 ROME 中的 causal trace

实验使用Qwen1.5-MoE-A2.7B-Chat，对每层的所有模块、MLP 层和 Attn 层进行 causal trace。
可以看出，结果与传统的 transformer 模型的 causal trace 有一定的相似性，同样可以体现 MLP 对事实的存储作用。

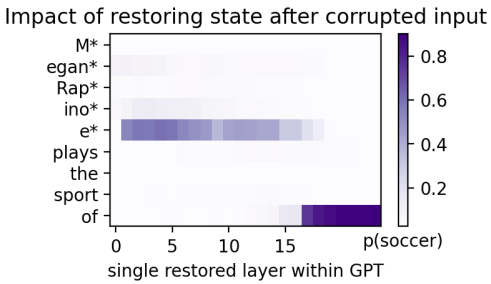


图 6: 全部层的 causal trace

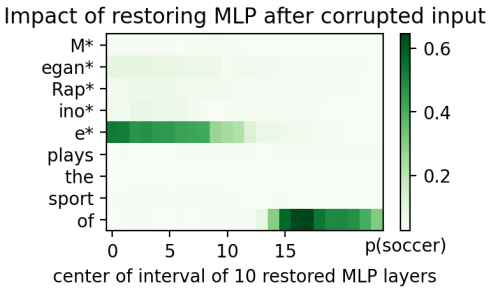


图 7: MLP 层的 causal trace

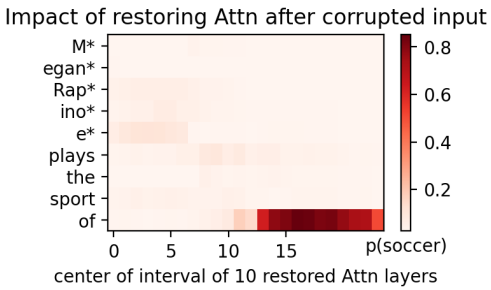


图 8: Attn 层的 causal trace

通过 causal trace 探究专家激活对输出的影响

通过破坏专家的参数后恢复某个专家，以此观察该专家对输出 token 的影响。

Qwen1.5-MoE 模型有 8 个层以及 64 个专家，其中每次激活 4 个专家。统计了每层的专家激活情况和专家路由情况。可以看出恢复某个专家对输出 token 的影响比较平均。

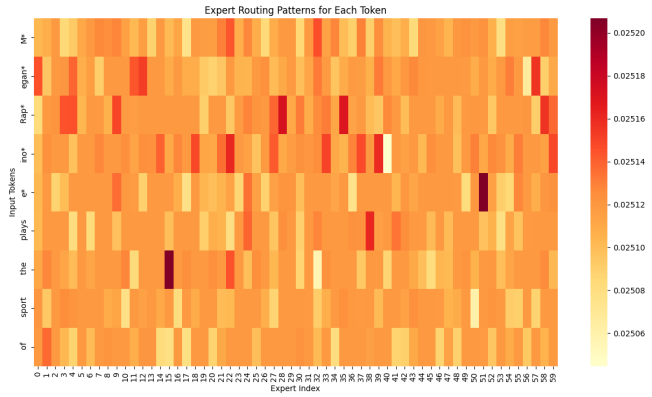


图 9: 专家路由情况

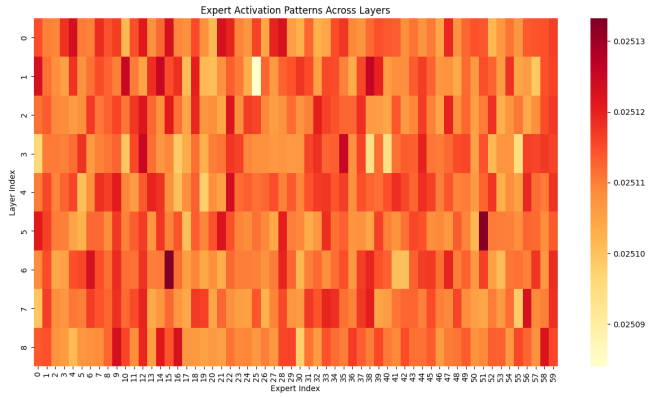


图 10: 专家激活情况

探究每 token 每层的主导专家

通过探究每个 token 每层的主导专家，可以发现每个 token 在不同层的主导专家是不同的，这也说明了 MoE 模型的专家激活是独立的。

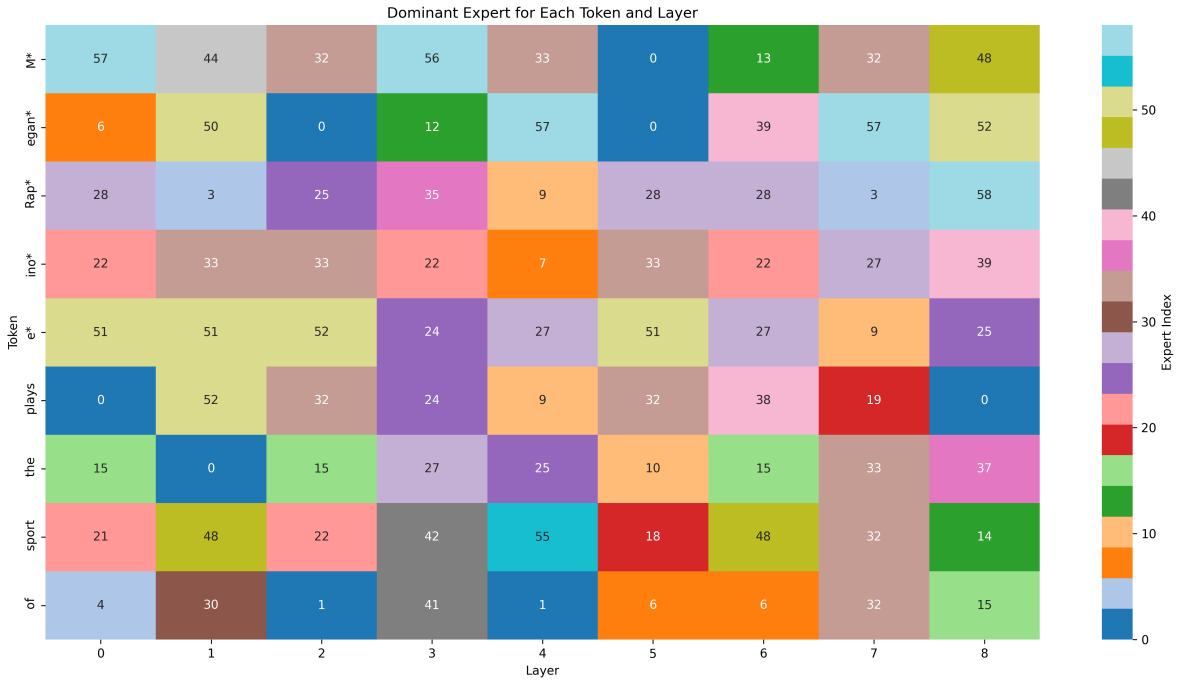


图 11: 每个 token 每层的主导专家

探究输入原始问题和复述问题的专家激活重合率

使用了 100 个事实进行测试，4 个专家中 3 个相同即算激活重合。结果发现激活重合率为 98%。

```
高一致性事实示例:
1. Atlanta Flames - 一致率: 100.00%
   原始问题: What is the final year of Atlanta Flames?
   复述问题: What year ended with Atlanta Flames?
   预测答案: 1980
   专家重合情况:
      层3: 重合数量=4/4 (100.00%)
        原问题专家: [33, 51, 10, 19]
        复述问题专家: [33, 51, 19, 10]
      层4: 重合数量=4/4 (100.00%)
        原问题专家: [16, 29, 41, 25]
        复述问题专家: [16, 29, 41, 25]
      层5: 重合数量=4/4 (100.00%)
        原问题专家: [22, 42, 9, 10]
        复述问题专家: [22, 42, 9, 10]
```

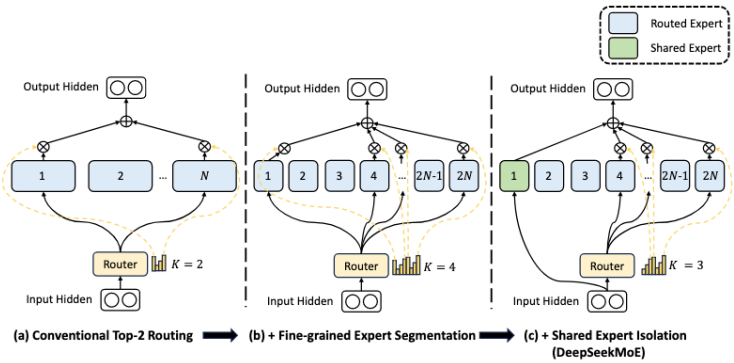
```
低一致性事实示例:
1. Didier Eribon - 一致率: 8.33%
   原始问题: What kind of  occupation does Didier Eribon have?
   复述问题: What kind of occupation did Didier Eribon have?
   预测答案: politician
   专家重合情况:
      层3: 重合数量=0/4 (0.00%)
        原问题专家: [43, 40, 47, 19]
        复述问题专家: [33, 51, 35, 10]
      层4: 重合数量=0/4 (0.00%)
        原问题专家: [4, 36, 56, 32]
        复述问题专家: [16, 29, 41, 25]
      层5: 重合数量=1/4 (25.00%)
        原问题专家: [3, 16, 32, 10]
        复述问题专家: [22, 42, 9, 10]
```

先定位专家再编辑

- 对中层/早期的 MLP 层，先进行激活专家的定位，再进行编辑
  - 使用类似 ROME 的求闭式解的方法
  - 也可以只对“主要专家”进行编辑（是否有助于提高泛化性？）
- 由于不同 token 激活专家的独立性，可能编辑过程较为复杂

编辑共享专家

一些 MoE 架构具有每次都会激活的“共享专家”，如 Qwen-MoE (4 合 1 共享专家 +60 个普通专家) 和 DeepSeek-MoE (1 共享专家 +63 普通专家)。
根据 DeepSeek-MoE 技术报告中的表述，共享专家中似乎保留着一些“常识”，是否可以尝试只编辑共享专家来更新知识？(或者与普通专家编辑相结合)



With a conventional routing strategy, tokens assigned to different experts may necessitate some common knowledge or information. As a result, multiple experts may converge in acquiring shared knowledge in their respective parameters, thereby resulting in redundancy in expert parameters. However, if there are shared experts dedicated to capturing and consolidating common knowledge across varying contexts, the parameter redundancy among other routed experts will be alleviated. This alleviation of redundancy will contribute to a more parameter-efficient model with more specialized experts.

References

- Efficiently Editing Mixture-of-Experts Models with Compressed Experts
- DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models
- Qwen1.5-MoE: Matching 7B Model Performance with 1/3 Activated Parameters
- Applying Mixture of Experts in LLM Architectures