Mixture of Experts (MoE)



Contents

1. 奠基工作: 90 年代初期

1991, Hitton, Adaptive Mixtures of Local Experts

2. 架构形成: RNN 时代

2017, Google, Outrageously Large Neural Networks

3. 提升效果: Transformer 时代

- 2020, Google, GShard
- 2022, Google, Switch Transformer

4. 智能涌现: GPT 时代

- 2021, Google, GLaM
- 。 2024, 幻方量化, DeepseekMoE/ Deepseek V2/ Deepseek V3





视频目录大纲

MoE 基本介绍 MoE 前世今生 奠基工作: Adaptive Mixtures of Local Experts LSTM MoE: Outrageously Large Neural Networks ... GShard: Scaling Giant Models with Conditional ... MoE 核心工作 Switch Transformers: Scaling to Trillion Parameter Models ... MoE 混合专家 GLaM: Efficient Scaling of Language Models with Mixture ... 好好想想// ST-MoE: Designing Stable and Transferable Sparse Ex MoE 架构原理 大模型遇上 MoE 手撕 MoE 代码 MoE 的未来

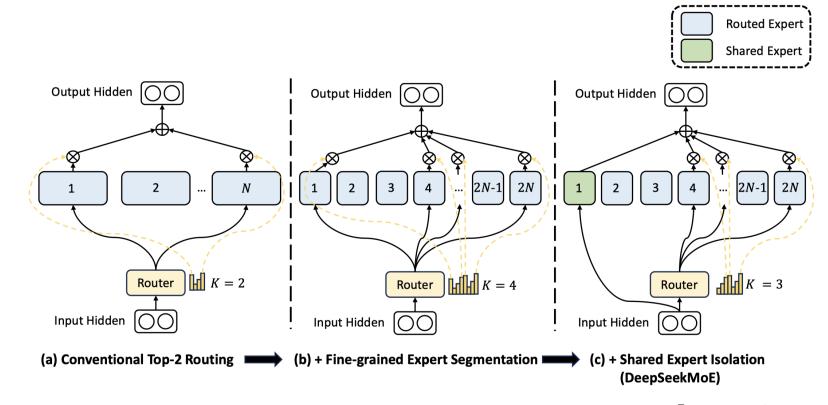


DeepSeek MOE



基本介绍

• DeepSeekMOE 是一种创新的混合专家(Mixture of Experts, MoE)模型架构,旨在通过细分割专家和共享专家策略,提高专家专业化程度,减少冗余,并在参数扩展时控制计算成本。其核心目标是在保持高性能的同时,显著降低计算资源消耗。





背景与动机

• 模型规模扩展的挑战:

随着模型参数量的增加,计算成本显著上升,限制了模型的扩展。DeepSeekMOE 通过稀疏专家模型 (MoE)解决这一问题,仅激活部分参数,从而在保持高性能的同时降低计算开销。

• 专家同质化与负载失衡:

传统 MoE 模型存在专家同质化和负载失衡问题,部分专家承担过多计算任务,而其他专家利用率低。DeepSeekMOE 通过动态专业化路由(DSR)和专家共享机制,优化专家利用率。



核心架构

细分割与共享专家策略:

细分割专家:将专家分为多个小组,每个小组专注于特定任务,提高专家专业化程度。

· **共享专家**:引入共享专家,捕捉通用知识,减少冗余参数。

• 动态专业化路由 (DSR) :

- DSR 通过门控网络动态选择最相关专家,确保每个输入仅激活少量专家,从而降低计算成本。
- DSR 还通过负载均衡机制,避免某些专家过载或闲置,确保专家利用率均衡

训练策略

• 三阶段训练流程:

。 专家孵化: 通过动态课程学习, 初步训练专家网络。

专精强化:通过对抗训练,进一步优化专家专业化能力。

协同优化:通过多任务协同训练,提升整体模型性能。

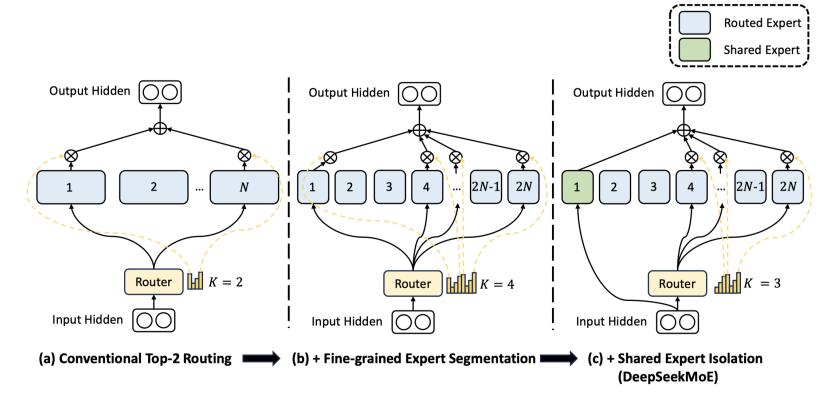
• 训练效率优化:

· 采用 FP8 混合精度训练和 DualPipe 算法,显著减少训练时间和通信开销



DeepSeek MOE

DeepSeekMOE 通过细分割专家、共享专家和动态专业化路由等创新技术,实现了高效的大规模模型训练和推理。其在保持高性能的同时,显著降低了计算成本,为万亿参数时代的大模型发展提供了新的思路。







把AI系统带入每个开发者、每个家庭、 每个组织,构建万物互联的智能世界

Bring Al System to every person, home and organization for a fully connected, intelligent world.

Copyright © 2024 XXX Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



 $Git Hub\ https://github.com/chenzomi \ I\ 2/Al \ Infra$

引用与参考

- https://mp.weixin.qq.com/s/6kzCMsJuavkZPG0YCKgeig
- https://www.zhihu.com/tardis/zm/art/677638939?source id=1003
- https://huggingface.co/blog/zh/moe
- https://mp.weixin.qq.com/s/mOrAYo3qEACjSwcRPG7fWw
- https://mp.weixin.qq.com/s/x39hqf8xn1cUlnxEIM0 ww
- https://mp.weixin.qq.com/s/ZXjwnO103e-wXJGmmKi-Pw
- https://mp.weixin.qq.com/s/8Y281VYaLu5jHoAvQVvVJg
- https://blog.csdn.net/weixin 43013480/article/details/139301000
- https://developer.nvidia.com/zh-cn/blog/applying-mixture-of-experts-in-llm-architectures/
- https://www.zair.top/post/mixture-of-experts/
- https://my.oschina.net/IDP/blog/16513157
- PPT 开源: https://github.com/chenzomi12/AlInfra
- 夸克链接: https://pan.quark.cn/s/74fb24be8eff

