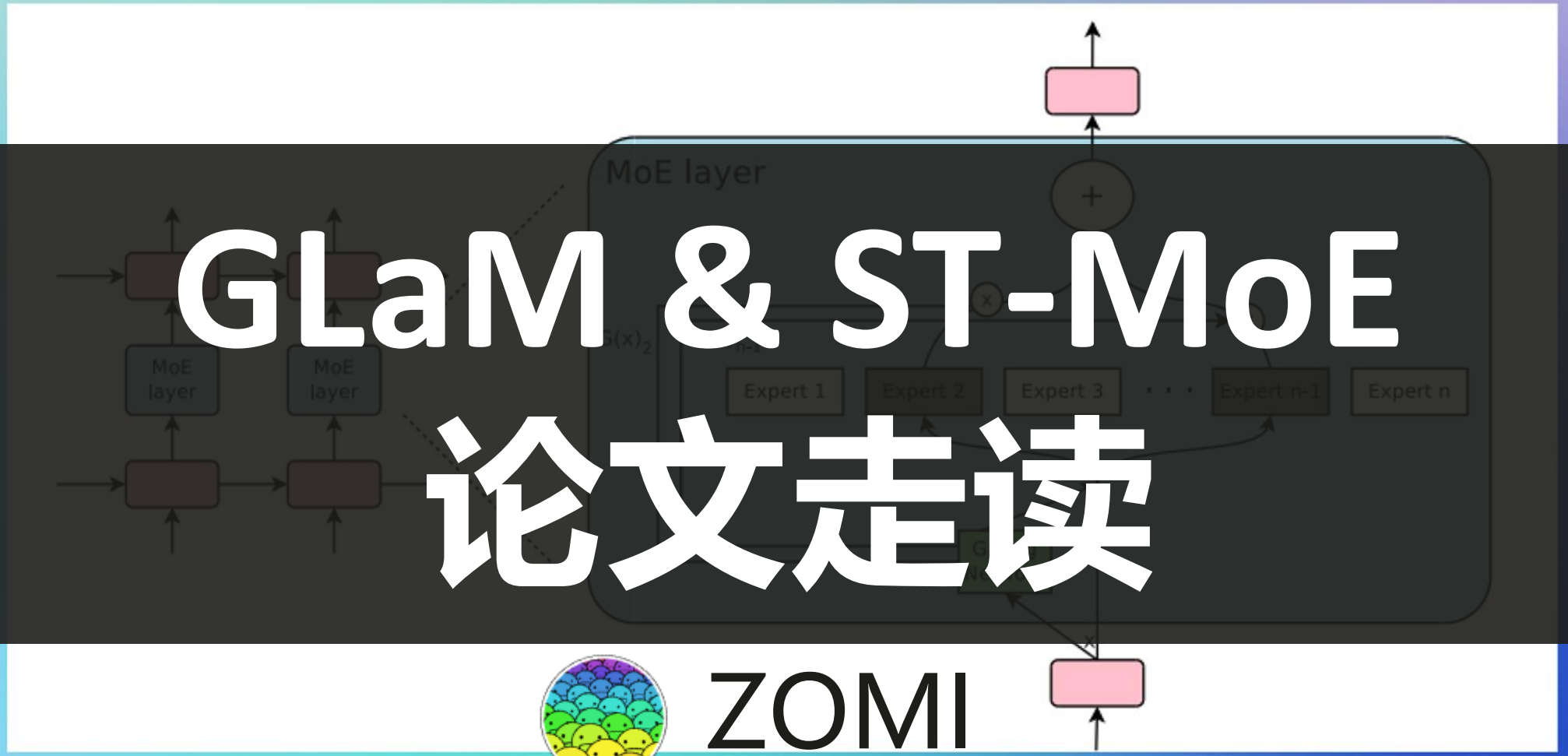
Mixture of Experts (MoE)



GLaM & ST-MoE 论文走读

ZOMI

Contents

视频目录大纲

- MoE 混合专家
  - MoE 基本介绍 (4)
  - MoE 前世今生 (5)
  - MoE 核心工作
    - 奠基工作：Adaptive Mixtures of Local Experts
    - LSTM MoE：Outrageously Large Neural Networks ...
    - GShard: Scaling Giant Models with Conditional ...
    - Switch Transformers: Scaling to Trillion Parameter Models ...
    - GLaM: Efficient Scaling of Language Models with Mixture ...
    - ST-MoE: Designing Stable and Transferable Sparse Exp
  - MoE 架构原理 (3)
  - 大模型遇上 MoE (4)
  - 手撕 MoE 代码 (5)
  - MoE 的未来 (2)

05

ST-MoE

ST-MOE 研究动机

- ST-MoE（Designing Stable and Transferable Sparse Expert Models）是谷歌团队提出的一种稀疏混合专家模型，专注于解决稀疏模型在训练稳定性和迁移学习中的问题。
- **研究动机**
  - ST-MoE 的研究动机在于解决稀疏模型在预训练阶段的不稳定性以及微调阶段的质量问题。稀疏模型虽然参数量大，但实际计算量较少，具有高效性，但其训练过程容易出现不稳定现象

ST-MOE

- **核心创新：Router z-loss**
  - 论文提出了一种新的辅助损失函数——Router z-loss，用于提高稀疏模型的训练稳定性。这种损失函数通过对路由决策进行约束，避免了某些专家被过度激活或完全未被使用的情况，从而提升了模型的鲁棒性
- **模型规模与性能：**
  - ST-MoE 通过将稀疏模型扩展到 2690 亿参数，展示了其在多种任务上的全面迁移能力。这是第一个能够在各类任务中达到最先进性能的稀疏模型 。
- **架构设计原则：**
  - 论文提出了设计稀疏模型的架构、路由和模型设计原则，尤其是在分布式环境中的效率优化。这些原则确保了模型在扩展参数规模的同时保持高效的计算性能 。

ST-MOE

- **负载均衡与专家分布**
  - ST-MoE 引入了噪声机制来改善门控网络的负载均衡，防止某些专家过载或闲置。此外，论文还分析了专家层中的 token 路由决策，揭示了不同专家对特定 token 组的专长分布
- **迁移学习能力**
  - ST-MoE 不仅在大规模预训练中表现出色，还在微调阶段展现了优秀的可迁移性。这使得稀疏模型能够更好地适应下游任务，解决了传统稀疏模型在迁移学习中的局限性 。
- **背景与相关工作**
  - ST-MoE 是在 Switch Transformers 和 Sparsely-Gated Mixture 等稀疏模型的基础上进一步发展的。它继承了稀疏混合专家模型（MoE）的基本思想，同时针对训练不稳定性和迁移性问题进行了改进

05

GLaM

基本介绍

- GLaM 论文提出了一种基于稀疏激活专家混合架构的通用语言模型，通过优化路由机制和负载均衡策略，在保持高性能的同时大幅降低了训练和推理的能耗。其在多任务学习和少样本学习中的优异表现，以及对超大规模模型扩展性的探索，为自然语言处理领域的发展提供了重要的理论和实践指导。

Thank you

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

ZOMI

GitHub https://github.com/chenzomi12/AIInfra

引用与参考

- https://zhuanlan.zhihu.com/p/653796685

- PPT 开源：https://github.com/chenzomi12/AIInfra
- 夸克链接：https://pan.quark.cn/s/74fb24be8eff