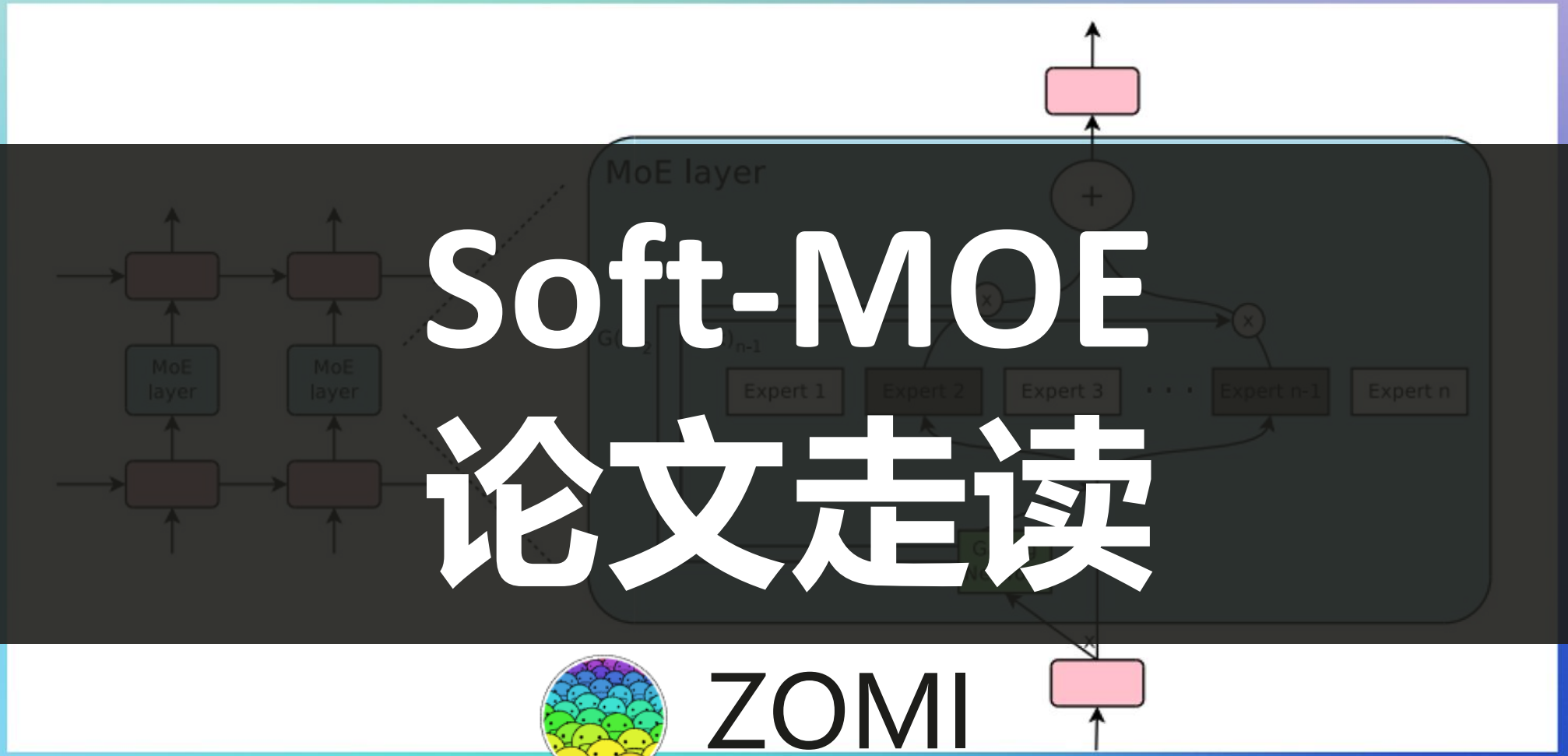
Mixture of Experts (MoE)



Soft-MOE
论文走读

ZOMI

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 6
  - MoE 架构原理 3
  - 大模型遇上 MoE 4
  - 手撕 MoE 代码 5
  - MoE 的未来
    - V-MoE：视觉和多任务 MoE
    - Soft MoE：可微的稀疏混合专家

Soft-MOE

V-MOE

- Soft-MoE引入软分配（soft assignment）机制，通过加权组合所有输入token生成专家输入，避免了硬分配带来的问题。具体来说，每个专家的输入是所有token的加权平均，权重由token和专家的相似性决定。

全可微分架构

- **连续性与可微性：**
  - Soft-MoE的所有操作（包括路由和专家处理）都是连续且可微分的，这使得模型能够通过标准反向传播进行端到端训练，而无需复杂的负载均衡策略或辅助损失函数。
- **避免token丢失：**
  - 由于所有token都参与每个专家的输入计算，Soft-MoE完全避免了token丢失问题，同时保持了专家负载的均衡。

高效计算与性能提升

- **低推理成本：**
  - Soft-MoE在推理时仅需处理加权组合的token子集，显著降低了计算成本。例如，Soft-MoE Base/16的推理成本比ViT Huge/14低10.5倍，同时性能相当。
- **性能优势：**
  - 在视觉识别任务中，Soft-MoE在少样本学习和微调任务上优于标准Transformer和流行的MoE变体（如Token Choice和Expert Choice）。

后续影响

- Soft-MoE为混合专家模型的设计提供了新的思路，其软分配机制和全可微分架构解决了传统MoE的诸多问题，同时保持了高效的计算性能和扩展能力。该方法的提出不仅推动了视觉Transformer的发展，也为其他领域（如自然语言处理和多模态学习）的模型设计提供了借鉴。

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://www.youtube.com/watch?v=sOPDGQjFcuM&t=1s
- https://arxiv.org/abs/2308.00951
- https://arxiv.org/abs/2106.05974
- https://zhuanlan.zhihu.com/p/652536107

- PPT 开源在：https://github.com/chenzomi12/AIInfra