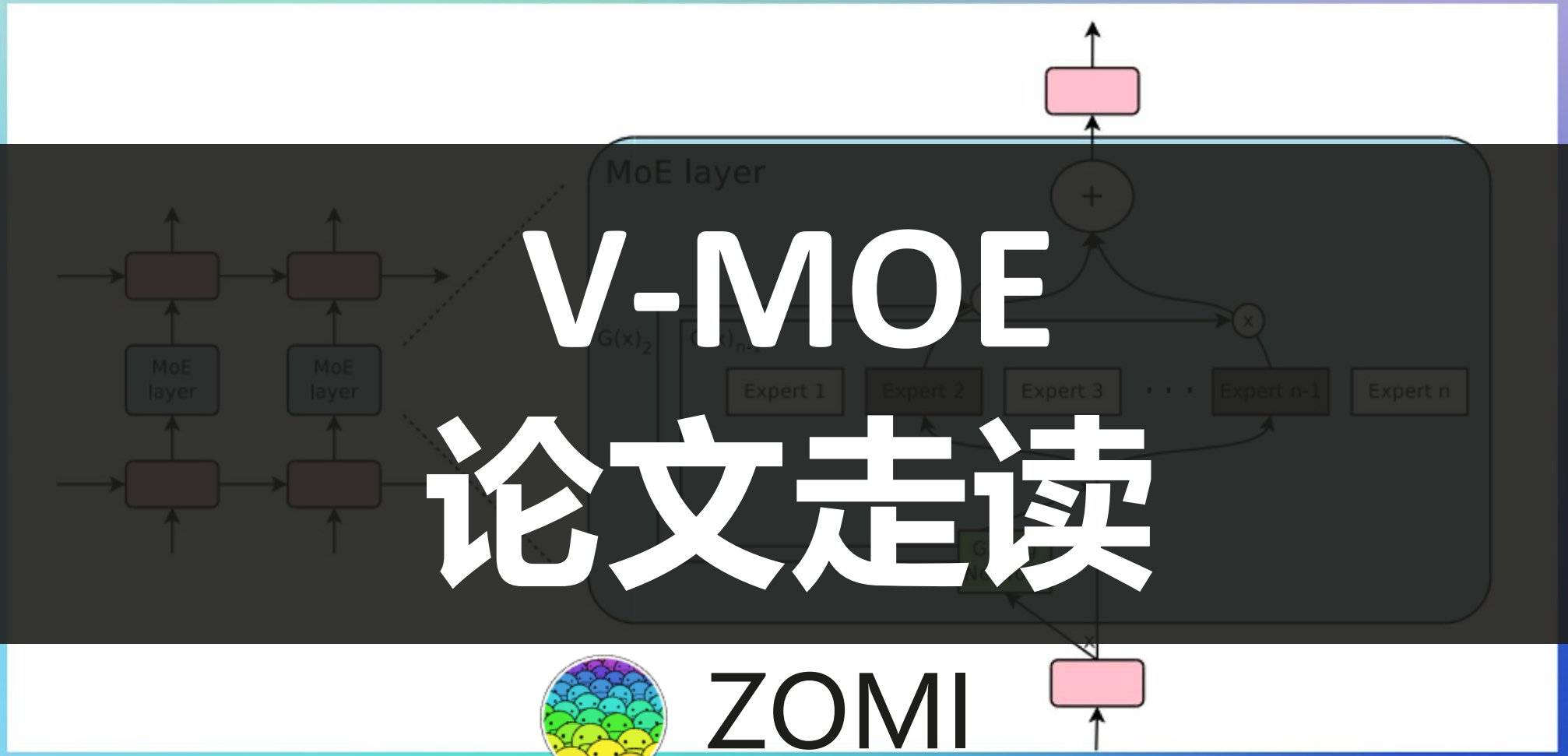
Mixture of Experts (MoE)



V-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

V-MOE

V-MOE

- 稀疏门控混合专家网络 (Sparsely-gated Mixture of Experts networks, MoE) 这种方法已经在自然语言处理领域中表现出了出色的可扩展性。但是在计算机视觉中，几乎所有性能网络都是 "密集 (Dense) 的"，也就是说，每个输入都由所有的参数来处理。

V-MOE

- 本文提出了视觉领域经典的稀疏门控混合专家网络 Vision MoE (V-MoE)，它是 Vision Transformer 的稀疏版本，V-MoE 是一种可扩展的架构，其性能和最大的密集网络适配。在图像识别任务中，V-MoE 与 SOTA 的网络的性能相匹配，同时在推理时只需要不到一半的 FLOPs。作者还对路由算法进行了扩展，对整个批次中每个输入的子集进行优先级排序，从而对每幅图像实现自适应的计算，使得测试时的计算更加平滑。
-

架构设计

- **稀疏MoE层替代稠密FFN：**
  - V-MoE将传统ViT中的前馈神经网络（FFN）层替换为稀疏混合专家层。每个MoE层包含多个专家（如8个），每个专家本身是一个独立的FFN，而门控网络（路由机制）动态决定输入图像块（token）分配给哪些专家处理。
- **动态路由与负载均衡：**
  - 门控网络通过Top-K选择机制（如选择1-2个专家）分配输入，并引入噪声和辅助损失（如变异系数优化）来均衡专家负载，避免某些专家被过度激活或训练不足。

后续影响

- V-MoE为视觉大模型的训练与部署提供了新范式，其设计思想被后续工作（如谷歌的Switch Transformer、Time-MoE时序模型）借鉴，推动了稀疏模型在CV、NLP等领域的应用1824。尽管存在显存占用和微调挑战，其高效扩展能力使其成为资源受限场景下训练大规模视觉模型的优选方案。

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