

TopoMoE：去中心化混合专家模型（修订版）

Lanhaijian 大模型辅助

虹桥大学科技学术网络

声明：虹桥大学实体机构暂未落成，虹桥大学科技学术网络已投入运行

1 摘要

本文引入宏观有效对偶耦合 C_{eff} 与自适应形变融合机制，对 TopoMoE 框架进行修订。TopoMoE 基于张量拓扑优化与离散外微分实现分层信息流约束，将非交换融合设计封装为独立宏观工程组件。本文明确： C_{eff} 与形变算子均为独立宏观工程参数，不依赖底层量子时空代数。全部模块与原始 TopoMoE 向后兼容，并在去中心化多专家协同任务上完成验证。

2 1 修订版 TopoMoE 框架

2.1 1.1 原始核心模块回顾

TopoMoE 是基于核化张量拓扑与离散外微分构建的去中心化混合专家（MoE）架构。模型将专家建模为 0-单纯形，专家交互边为 1-单纯形，三角协同群组为 2-单纯形。四类分层拓扑约束保证信息流守恒：梯度损失 $\mathcal{L}_{\text{grad}}$ 、散度损失 $\mathcal{L}_{\text{div}}$ 、旋度损失 $\mathcal{L}_{\text{curl}}$ 、曲率损失 $\mathcal{L}_{\text{curv}}$ 。原始 TopoMoE 总损失为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{grad}} + \lambda_1 \mathcal{L}_{\text{div}} + \lambda_2 \mathcal{L}_{\text{curl}} + \lambda_3 \mathcal{L}_{\text{curv}}. \quad (1)$$

2.2 1.2 专家宏观有效对偶耦合

去中心化 MoE 系统中，专家间信息传播分为两种典型状态：弱退相干交互、强相干闭环交互。本文引入宏观有效对偶耦合 $C_{\text{eff}}^{(e)}$ 动态区分两种状态并切换融合策略。

2.2.1 $C_{\text{eff}}^{(e)}$ 定义

对专家 e ，记 $\tau_{\text{coh}}^{(e)}$ 为相干推理步数， $\tau_{\text{delay}}^{(e)}$ 为专家间平均消息延迟步数， W_{eh} 为专家 e 与邻域专家 h 的拓扑权重， $\text{Var}(r_h)$ 为邻域专家 h 输出方差， $\sigma^{\mathcal{N}(e)}$ 为邻域输出标准差，数值稳定项 $\epsilon = 10^{-8}$ 。归一化宏观有效对偶耦合定义为：

$$\mathcal{C}_{\text{eff}}^{(e)} = \frac{\tau_{\text{coh}}^{(e)}}{\tau_{\text{delay}}^{(e)}} \cdot \frac{1}{|\mathcal{N}(e)|} \sum_{h \in \mathcal{N}(e)} W_{eh} \cdot \frac{\text{Var}(r_h)}{(\sigma^{\mathcal{N}(e)})^2 + \epsilon}. \quad (2)$$

2.2.2 临界阈值与运行区间

记 C_c 为耦合临界阈值，可设为固定超参或可学习参数，划分两类运行区间：

- $\mathcal{C}_{\text{eff}}^{(e)} < C_c$ ：弱耦合区间，采用经典张量拼接 + 多层感知机融合；

- $\mathcal{C}_{\text{eff}}^{(e)} \geq C_c$: 强相干区间, 启用形变融合机制。

2.3 1.3 自适应张量消息融合

设 x_e 为专家 e 的张量状态, msg_e 为聚合后的专家间张量消息, 自适应融合规则:

$$\tilde{x}_e = \begin{cases} \text{Concat}(x_e, \sigma(\text{MLP}[x_e; \text{msg}_e]) \odot \text{msg}_e) & \mathcal{C}_{\text{eff}}^{(e)} < \mathcal{C}_c \\ \text{Def-Fuse}(x_e, \text{msg}_e; \tilde{\theta}) & \mathcal{C}_{\text{eff}}^{(e)} \geq \mathcal{C}_c \end{cases} \quad (3)$$

2.4 1.4 Def-Fuse: 形变非交换融合算子

定义可学习线性映射 $\phi(\cdot) = W_\phi \cdot + b_\phi$, $\psi(\cdot) = W_\psi \cdot + b_\psi$, 形变融合算子表达式:

$$\begin{aligned} \text{Def-Fuse}(x_e, \text{msg}_e; \tilde{\theta}) &= W_1 x_e + W_2 \text{msg}_e \\ &\quad + \cos(2\pi\tilde{\theta}) \cdot (\phi(x_e) \odot \psi(\text{msg}_e) - \phi(\text{msg}_e) \odot \psi(x_e)). \end{aligned} \quad (4)$$

2.4.1 非交换性验证

交换输入后算子差值不为零, 证明非交换特性:

$$\begin{aligned} &\text{Def-Fuse}(\text{msg}_e, x_e) - \text{Def-Fuse}(x_e, \text{msg}_e) \\ &= 2 \cos(2\pi\tilde{\theta}) (\phi(\text{msg}_e) \odot \psi(x_e) - \phi(x_e) \odot \psi(\text{msg}_e)) \neq 0. \end{aligned} \quad (5)$$

当 $\tilde{\theta} \rightarrow 0$ 时, 非交换项消失, Def-Fuse 退化为经典线性融合。

2.5 1.5 动态拓扑约束权重

将原始固定权重扩展为随 $\mathcal{C}_{\text{eff}}$ 自适应调整的动态权重:

$$\begin{cases} \lambda_1 = \lambda_1^0, \lambda_2 = \lambda_2^0, \lambda_3 = \lambda_3^0, & \mathcal{C}_{\text{eff}}^{(e)} < \mathcal{C}_c \\ \lambda_1 = \lambda_1^0, \lambda_2 = \lambda_2^0 \cdot (1 + \gamma \cdot \mathcal{C}_{\text{eff}}^{(e)}), \lambda_3 = \lambda_3^0 \cdot (1 + \gamma \cdot \mathcal{C}_{\text{eff}}^{(e)}), & \mathcal{C}_{\text{eff}}^{(e)} \geq \mathcal{C}_c \end{cases} \quad (6)$$

其中缩放因子 $\gamma > 0$ 。强相干区间下旋度、曲率约束权重提升, 抑制循环逻辑冲突。

2.6 1.6 更新后总损失函数

最终总损失表达式:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{grad}} + \lambda_1^0 \mathcal{L}_{\text{div}} + \lambda_2(\mathcal{C}_{\text{eff}}) \mathcal{L}_{\text{curl}} + \lambda_3(\mathcal{C}_{\text{eff}}) \mathcal{L}_{\text{curv}}. \quad (7)$$

3 2 训练算法

4 3 兼容性说明

修订版 TopoMoE 与原始框架完全向后兼容。当 $\mathcal{C}_{\text{eff}}^{(e)} \rightarrow 0$ 或 $\tilde{\theta} \rightarrow 0$ 时, 自适应融合机制与动态权重自动退化至原始固定参数框架。

Algorithm 1 修订版 TopoMoE 训练算法

- 1: 初始化专家网络、张量编码器与去中心化消息交互模块
 - 2: 初始化超参数: $\lambda_1^0, \lambda_2^0, \lambda_3^0, \gamma, \eta_{\tilde{\theta}}$
 - 3: 初始化临界阈值 C_c 、初始形变参数 $\tilde{\theta}(0)$
 - 4: **while** 训练未收敛 **do**
 - 5: 构建单纯复形: 为全部专家构造 0-1-2 单纯形拓扑结构
 - 6: 采集推理轨迹与张量特征
 - 7: **for** 每一个专家 e **do**
 - 8: 计算 $\tau_{\text{coh}}^{(e)}, \tau_{\text{delay}}^{(e)}, \text{Var}(r_h), \sigma^{\mathcal{N}(e)}$
 - 9: 依据式 (2) 计算 $\mathcal{C}_{\text{eff}}^{(e)}$
 - 10: 聚合得到邻域张量消息 msg_e
 - 11: 依据式 (3) 计算融合张量状态 $\tilde{x}_e$
 - 12: 计算张量优势项 A_{KT}^e
 - 13: 依据式 (6) 更新动态约束权重 λ_2, λ_3
 - 14: **end for**
 - 15: 依据式 (7) 计算总损失 $\mathcal{L}_{\text{total}}$
 - 16: 反向传播更新全部专家网络与编码器参数
 - 17: **if** 存在专家满足 $\mathcal{C}_{\text{eff}}^{(e)} \geq C_c$ **then**
 - 18: 梯度下降更新形变参数 $\tilde{\theta}$
 - 19: **end if**
 - 20: **end while**
 - 21: **return** 全部专家网络与张量编码器
-