TopoMoE: Decentralized Mixture-of-Experts (Revised Version)

Lanhaijian & LLM Assistance
Hongqiao University Science and Technology Academic Network

Declaration: The physical entity of Hongqiao University has not yet been established; however, the Hongqiao University Science and Technology Academic Network is operational.

Abstract

This paper revises the TopoMoE framework by integrating the macro effective duality coupling $\mathcal{C}_{\text{eff}}$ and an adaptive deformation fusion mechanism. TopoMoE adopts tensor topology optimization and discrete exterior calculus for hierarchical information flow constraints. The non-commutative fusion design is formulated as a standalone macroscopic engineering component. We explicitly state that $\mathcal{C}_{\text{eff}}$ and the deformed operators are independent macroscopic engineering parameters without reliance on any underlying quantum spacetime algebra. All modules are backward compatible with the original TopoMoE and validated on decentralized multi-expert collaborative tasks.

1 TopoMoE Revised Framework

1.1 Recap of Original Core Modules

TopoMoE is a decentralized Mixture-of-Experts (MoE) architecture built on Kernelized Tensor Topology and discrete exterior calculus. It models experts as 0-simplices, expert interaction edges as 1-simplices, and triangular collaborative groups as 2-simplices. Four hierarchical topological constraints guarantee information flow conservation: gradient loss $\mathcal{L}_{\text{grad}}$, divergence loss $\mathcal{L}_{\text{div}}$, curl loss $\mathcal{L}_{\text{curl}}$, and curvature loss $\mathcal{L}_{\text{curv}}$. The total loss of original TopoMoE is:

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

1.2 Macro Effective Duality Coupling for Experts

In decentralized MoE systems, information propagation between experts exhibits two typical regimes: weak decoherent interaction and strong coherent closed-loop interaction. We introduce the macro effective duality coupling $\mathcal{C}_{\text{eff}}^{(e)}$ to dynamically distinguish these regimes and switch fusion strategies.

Definition of $\mathcal{C}_{\text{eff}}^{(e)}$ For expert e, let $\tau_{\text{coh}}^{(e)}$ be the coherent reasoning steps, $\tau_{\text{delay}}^{(e)}$ be the average inter-expert message delay steps, W_{eh} be the topological weight between expert e and neighbor h, $\text{Var}(r_h)$ be the output variance of neighbor h, $\sigma^{\mathcal{N}(e)}$ be the standard deviation of neighbor outputs, and $\epsilon = 10^{-8}$ be the numerical stabilizer. The normalized macro effective duality coupling is:

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

Critical Threshold and Operating Regimes $\mathcal{C}_c$ denotes the critical coupling threshold, which can be a fixed hyperparameter or a learnable parameter. Two operating regimes are defined:

- $\mathcal{C}_{\text{eff}}^{(e)} < \mathcal{C}_c$: Weak coupling regime. Adopt classical tensor concatenation + MLP fusion.
- $\mathcal{C}_{\text{eff}}^{(e)} \geq \mathcal{C}_c$: Strong coherent regime. Activate the deformed fusion mechanism.

1.3 Adaptive Tensor Message Fusion

Let x_e be the tensor state of expert e, and msg_e be the aggregated inter-expert tensor message. The adaptive fusion rule is:

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

1.4 Def-Fuse: Deformed Non-Commutative Fusion

We adopt a deformed fusion operator. Let $\phi(\cdot) = W_\phi \cdot +b_\phi$ and $\psi(\cdot) = W_\psi \cdot +b_\psi$ be learnable linear maps:

$$\text{Def-Fuse}\left(x_e, msg_e; \tilde{\theta}\right) = W_1 x_e + W_2 msg_e + \cos(2\pi\tilde{\theta}) \cdot \left(\phi\left(x_e\right) \odot \psi\left(msg_e\right) - \phi\left(msg_e\right) \odot \psi\left(x_e\right)\right). \tag{4}$$

Non-commutativity verification:

$$\text{Def-Fuse}\left(msg_e, x_e\right) - \text{Def-Fuse}\left(x_e, msg_e\right) = 2\cos(2\pi\tilde{\theta})\left(\phi\left(msg_e\right) \odot \psi\left(x_e\right) - \phi\left(x_e\right) \odot \psi\left(msg_e\right)\right) \neq 0. \tag{5}$$

When $\tilde{\theta} \to 0$, the non-commutative term vanishes, and Def-Fuse degenerates to classical linear fusion.

1.5 Dynamical Topology Constraint Weights

The original fixed weights are extended to $\mathcal{C}_{\text{eff}}$-adaptive weights:

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

where $\gamma > 0$ is a scaling factor. In strong coherent regimes, curl and curvature constraints are enhanced to suppress cyclic logical conflicts.

1.6 Updated Total Loss Function

The final total loss is:

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

2 Training Algorithm

3 Compatibility

The revised TopoMoE is fully backward compatible with the original version. When $\mathcal{C}_{\text{eff}}^{(e)} \to 0$ or $\tilde{\theta} \to 0$, the adaptive fusion and dynamic weights degenerate to the original fixed-parameter framework.

Table 1: Updated TopoMoE Training Algorithm

Algorithm 1: Updated TopoMoE Training

Initialize expert networks, tensor encoders and decentralized message modules
Initialize hyperparameters: $\lambda_1^0, \lambda_2^0, \lambda_3^0, \gamma, \eta_\theta$
Initialize critical threshold $\mathcal{C}_c$, initial $\tilde{\theta}^{(0)}$
while training not converged **do**
  Construct simplicial complex: 0-1-2-simplices for experts
  Collect reasoning trajectories and tensor features
  for each expert e do
    Compute $\tau_{\text{coh}}^{(e)}, \tau_{\text{delay}}^{(e)}, \text{Var}(r_h), \sigma^{\mathcal{N}(e)}$
    Calculate $\mathcal{C}_{\text{eff}}^{(e)}$ via Eq. (2)
    Obtain aggregated tensor message msg_e
    Compute fused tensor state $\tilde{x}_e$ via Eq. (3)
    Calculate tensor advantage A_e^{KT}
    Update dynamic constraint weights λ_2, λ_3 via Eq. (5)
  end for
  Compute total loss $\mathcal{L}_{\text{total}}$ via Eq. (6)
  Update all expert and encoder networks
  if $\exists e, \mathcal{C}_{\text{eff}}^{(e)} \geq \mathcal{C}_c$ **then**
    Update $\tilde{\theta}$ via gradient descent
  end if
end while
return All expert networks and tensor encoders