Enhancing Mixture of Experts with Independent and Collaborative Learning for Long-Tail Visual Recognition

IJCAI-2025

Background&Architecture

1.专家同质化：所有专家倾向于学习相似的、偏向头部类别的特征，缺乏多样性。

2.优化目标冲突：简单地将所有专家的损失求和，可能导致专家间优化方向不一致，陷入次优解。

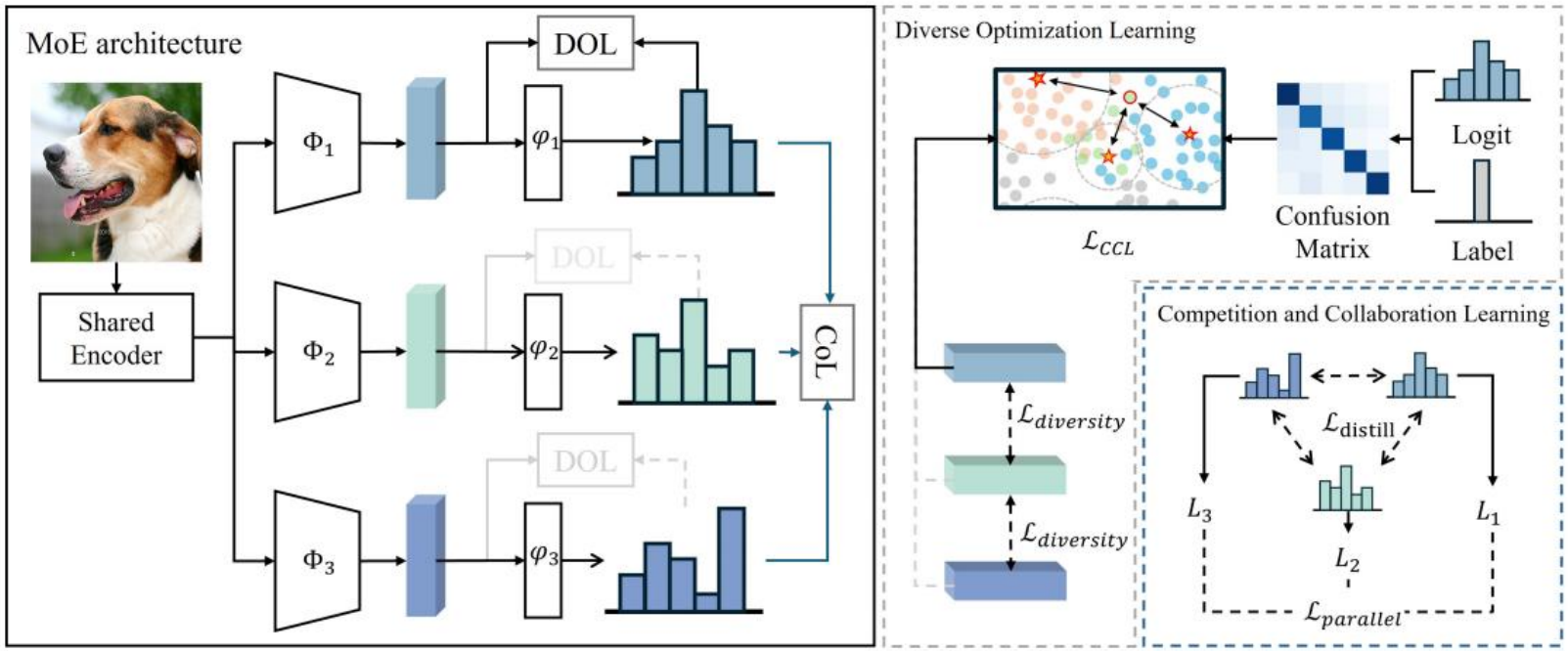


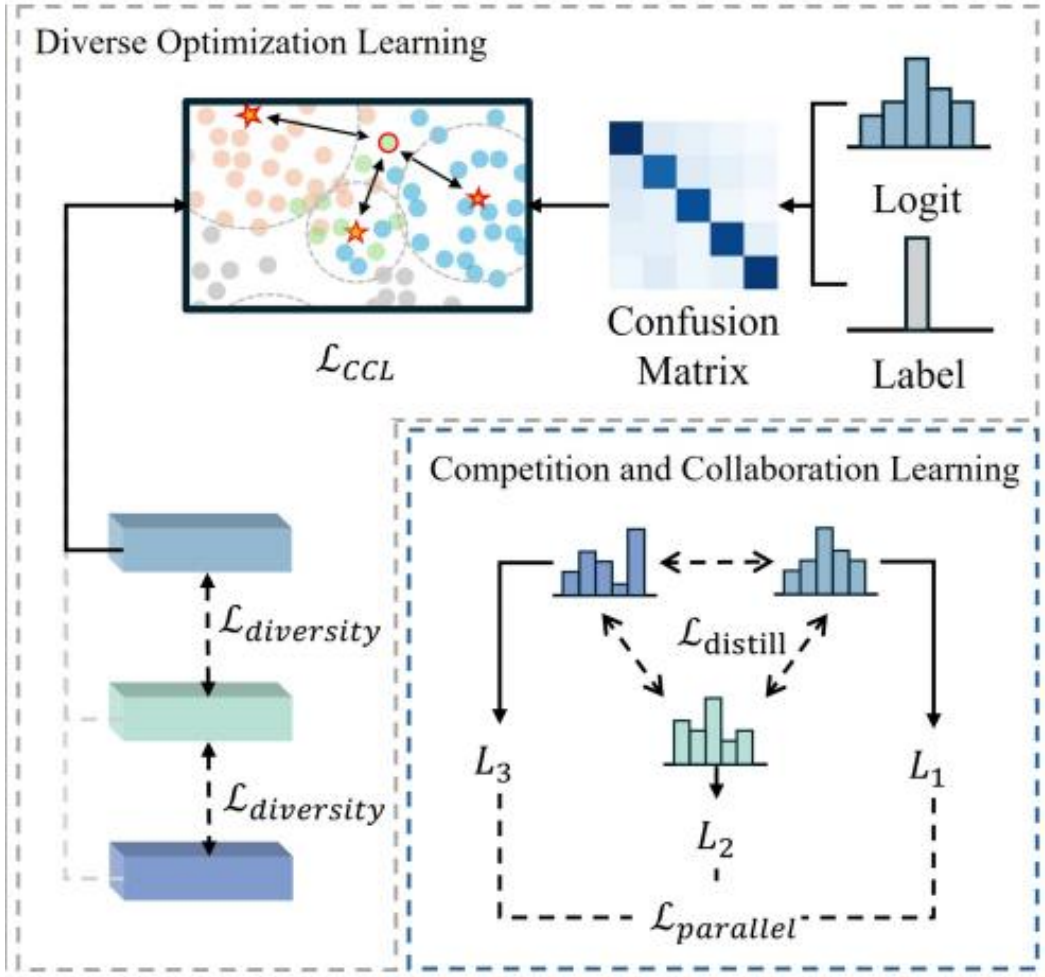
Figure 2: The proposed ICL contains two modules: DOL and CoL. Each expert in the MoE performs diversity optimization and confusion contrastive learning in the feature space and then applies competition and collaboration learning at the logits level for better model optimization from the individual to the overall.

作者提出了 ICL 框架，它包含两个核心的创新模块：

•DOL：多样化优化学习，作用于特征空间，主要解决专家同质化问题。

•CoL：竞争与协作学习，作用于Logits层面，主要解决优化冲突问题。

Diverse Optimization Learning



DOL（多样优化学习）：
•在特征空间中引入对比学习和混淆矩阵，强制专家多样化。
•解决传统 MoE 专家“同质化”的问题。

Adaptive Diversity (AD)

$$\mathcal{L}_{\text{AD}} = \sum_{m=1}^{M} \sum_{m'=1}^{M} \mathbb{I}_{m \neq m'} \frac{\mathbf{h}^m \cdot \mathbf{h}^{m'}}{\|\mathbf{h}^m\| \|\mathbf{h}^{m'}\|}$$

Confusion Contrastive Learning (CCL)

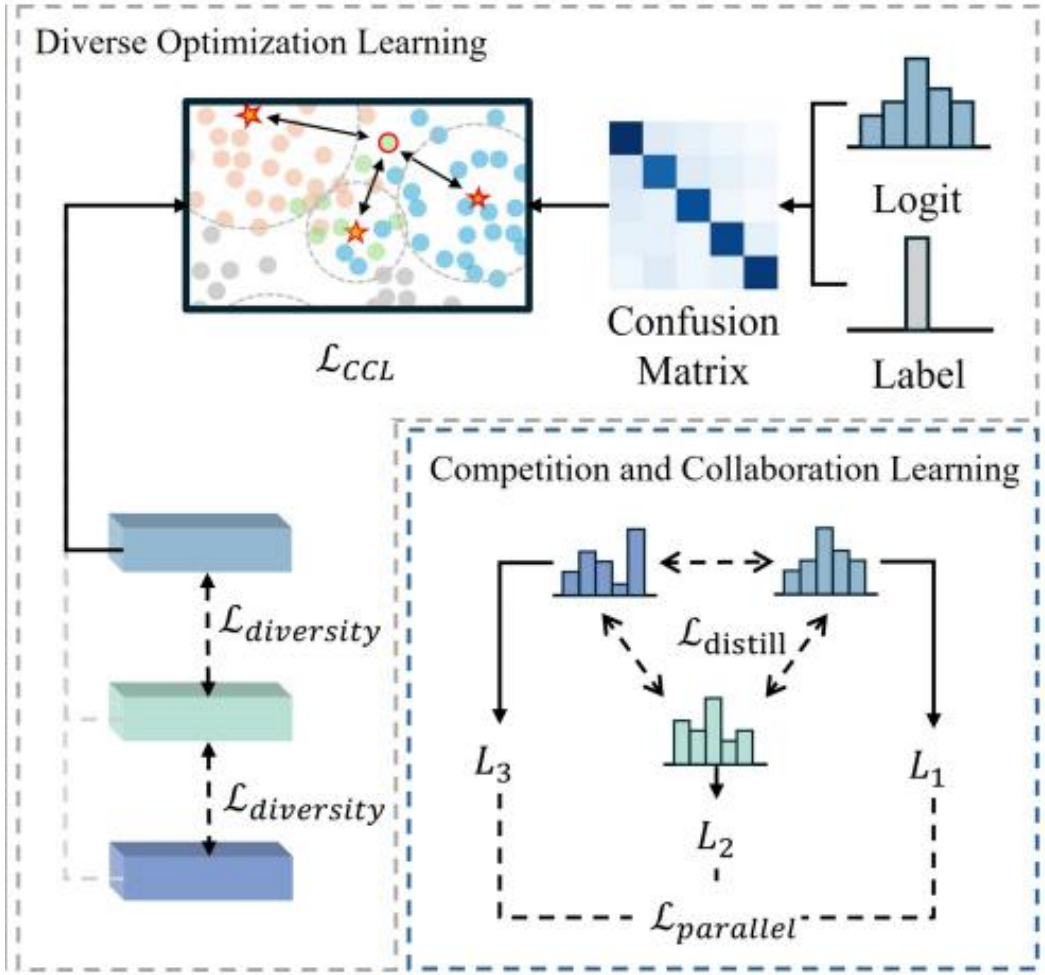
$$\mathcal{L}_{\text{CCL}} = -\log \frac{\exp\left(\frac{\mathbf{h}_i^m \cdot \mathbf{p}_c^m}{\tau}\right)}{\exp\left(\frac{\mathbf{h}_i^m \cdot \mathbf{p}_c^m}{\tau}\right) + \sum_{c' \in \mathcal{N}(c)} \exp\left(\frac{\mathbf{h}_i^m \cdot \mathbf{p}_{c'}^m}{\tau}\right)}$$

$\mathbf{h}_c^m$ is the mean feature vector of class c in the current batch for expert m.

DOL loss

$$\mathcal{L}_{\text{DOL}} = \mathbb{I}_{t<t_{AD}} \mathcal{L}_{\text{AD}} + \mathbb{I}_{t>t_{CL}} \sum_{m=1}^{M} \mathcal{L}_{\text{CCL}}$$

Competition and Collaboration Learning



CoL（竞争与协作学习）：

•在Logits 层面引入竞争机制（梯度放大）和协作机制（相互蒸馏）。

•解决传统 MoE 专家“目标冲突”和“局部最优”的问题。

Competition Learning（并行损失）

$$L_{\text{parallel}} = \left(\sum_{i=1}^{M} \frac{1}{L_i + \epsilon} \right)^{-1} \qquad \frac{\partial L_{\text{parallel}}}{\partial L_i} = \frac{1}{\left(\sum_{j=1}^{M} \frac{1}{L_j + \epsilon} \right)^2 (L_i + \epsilon)^2}$$

它的灵感来自于并联电路——在电路中，电阻越小，分得的电流越大。在这里，损失越小的专家，被认为在当前样本上表现越好，因此在反向传播中应该获得更大的梯度。

Collaboration Learning

$$\mathcal{L}_{\text{distill}} = \frac{2}{M(M-1)} \sum_{m<m'} \text{KL}(\mathbf{z}^m \parallel \mathbf{z}^{m'})$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CoL}} + \mathcal{L}_{\text{DOL}}$$

Experiments

Category	Method	CIFAR-100-LT			CIFAR-10-LT		
		100	50	10	100	50	10
Baseline	Cross Entropy	38.3	43.9	55.7	70.4	74.8	86.4
	Focal loss [Lin *et al.*, 2017]	38.7	46.2	–	74.6	79.3	–
Representation Learning	TSC [Li *et al.*, 2022d]	43.8	47.4	59.5	79.7	82.9	88.7
	BCL [Zhu *et al.*, 2022]	51.9	56.6	64.9	84.3	87.2	91.1
	SBCL [Hou *et al.*, 2023]	44.9	48.7	57.9	–	–	–
Re-balance	WD [Alshammari *et al.*, 2022]	52.4	57.4	67.9	–	–	–
	GCL [Li *et al.*, 2022c]	48.7	53.6	–	82.7	85.5	–
	KPS [Li *et al.*, 2023]	45.0	49.2	–	81.2	84.6	–
Data Augmentation	RISDA [Chen *et al.*, 2022]	50.2	53.8	62.4	79.9	84.2	89.4
	H2T [Li *et al.*, 2024]	51.4	55.5	–	–	–	–
	DBN-Mix [Baik *et al.*, 2024]	51.0	54.9	65.0	83.5	86.8	90.9
MoE	RIDE (3E) [Wang *et al.*, 2021]	49.1	–	–	–	–	–
	ACE (3E) [Cai *et al.*, 2021]	49.6	51.9	–	81.4	84.9	–
	TLC (4E) [Li *et al.*, 2022a]	49.8	–	–	80.4	–	–
	ResLT (3E) [Cui *et al.*, 2023]	49.7	54.5	63.7	–	–	–
	NCL (3E) [Li *et al.*, 2022b]	54.2	58.2	–	85.5	87.3	–
	SHIKE (3E) [Jin *et al.*, 2023]	56.3	59.8	–	–	–	–
	NCL++ (2E) [Tan *et al.*, 2024]	56.3	59.8	–	87.2	88.8	–
	Ours (2E)	56.3	59.7	69.0	86.4	88.5	91.3
	Ours (3E)	**57.6**	**61.3**	**69.3**	**87.9**	**89.7**	**91.9**

Method	Img-LT	iNat-LT
Single Model		
Cross Entropy	41.6	66.9
WD [Alshammari *et al.*, 2022]	53.3	70.0
BCL [Zhu *et al.*, 2022]	56.0	–
SBCL [Hou *et al.*, 2023]	57.1	70.8
GCL [Li *et al.*, 2022c]	54.9	–
GLMC [Du *et al.*, 2023]	56.3	–
H2T [Li *et al.*, 2024]	56.9	72.0
DBN-Mix [Baik *et al.*, 2024]	56.6	74.7
MoE-based Method		
RIDE [Wang *et al.*, 2021]	55.4	71.7
ACE [Cai *et al.*, 2021]	55.1	72.9
TLC [Li *et al.*, 2022a]	55.1	–
ResLT [Cui *et al.*, 2023]	55.1	72.9
NCL [Li *et al.*, 2022b]	59.5	74.9
SHIKE [Jin *et al.*, 2023]	59.7	75.4
NCL++ [Tan *et al.*, 2024]	59.6	75.2
Ours(2 experts)	59.5	75.3
Ours(3 experts)	**60.2**	**75.9**

Experiments

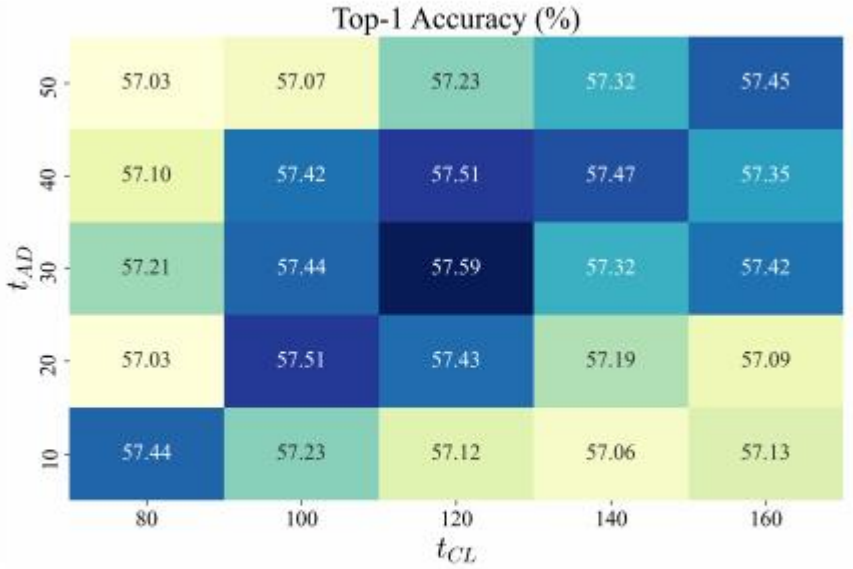
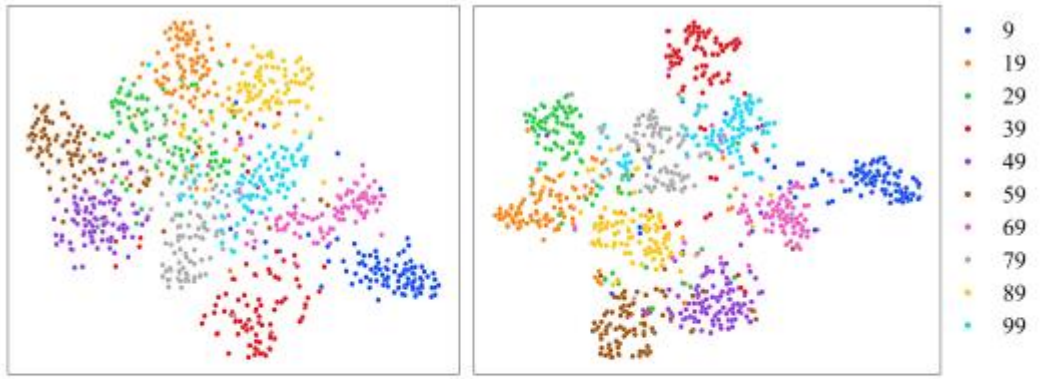
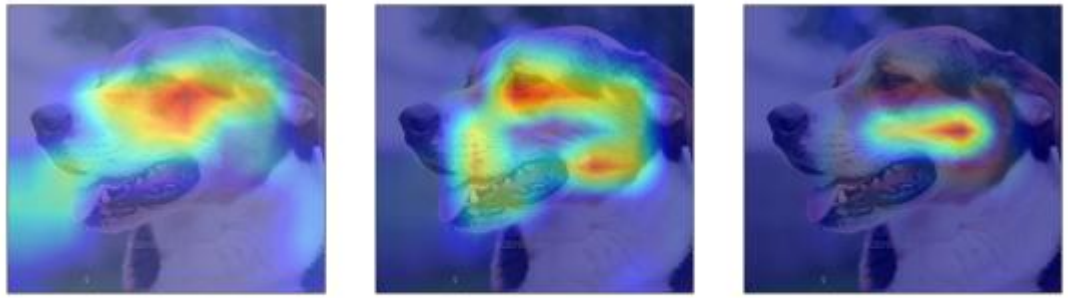


Figure 4: Parametric analysis of t_{AD} and t_{CL}.



(a) (b)

Figure 5: Visualization of t-SNE with (a) CE and (b) CCL on CIFAR100-LT with an imbalance factor of 100. We have chosen ten classes equally spaced for better view.



(a) Expert 1 (b) Expert 2 (c) Expert 3

Figure 6: Grad-CAM visualization.

Thanks