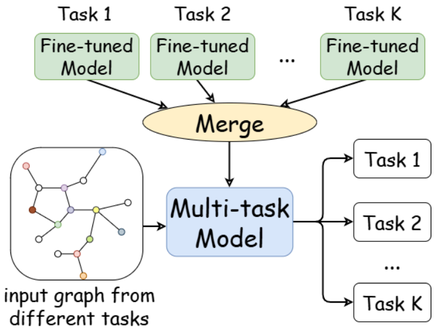
G-Merging: Graph Models Merging for Parameter-Efficient Multi-Task Knowledge Consolidation

Jun Chen [1 2], Ziyue Qiao [1], Qin Zhang [2], Kaize Ding [3], Xiao Luo [4]

[1] *Great Bay University* [2] *Shenzhen University* [3] *Northwestern University* [4] *University of Wisconsin-Madison*

ICLR International Conference On Learning Representations

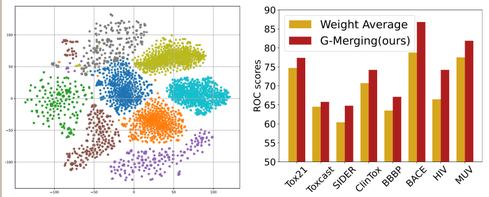
Problem Setup



Given: a pre-trained GNN and K task-specific fine-tuned GNNs.

Goal: **one unified multi-task graph model** that preserves task knowledge, reducing storage and deployment cost.

Challenges

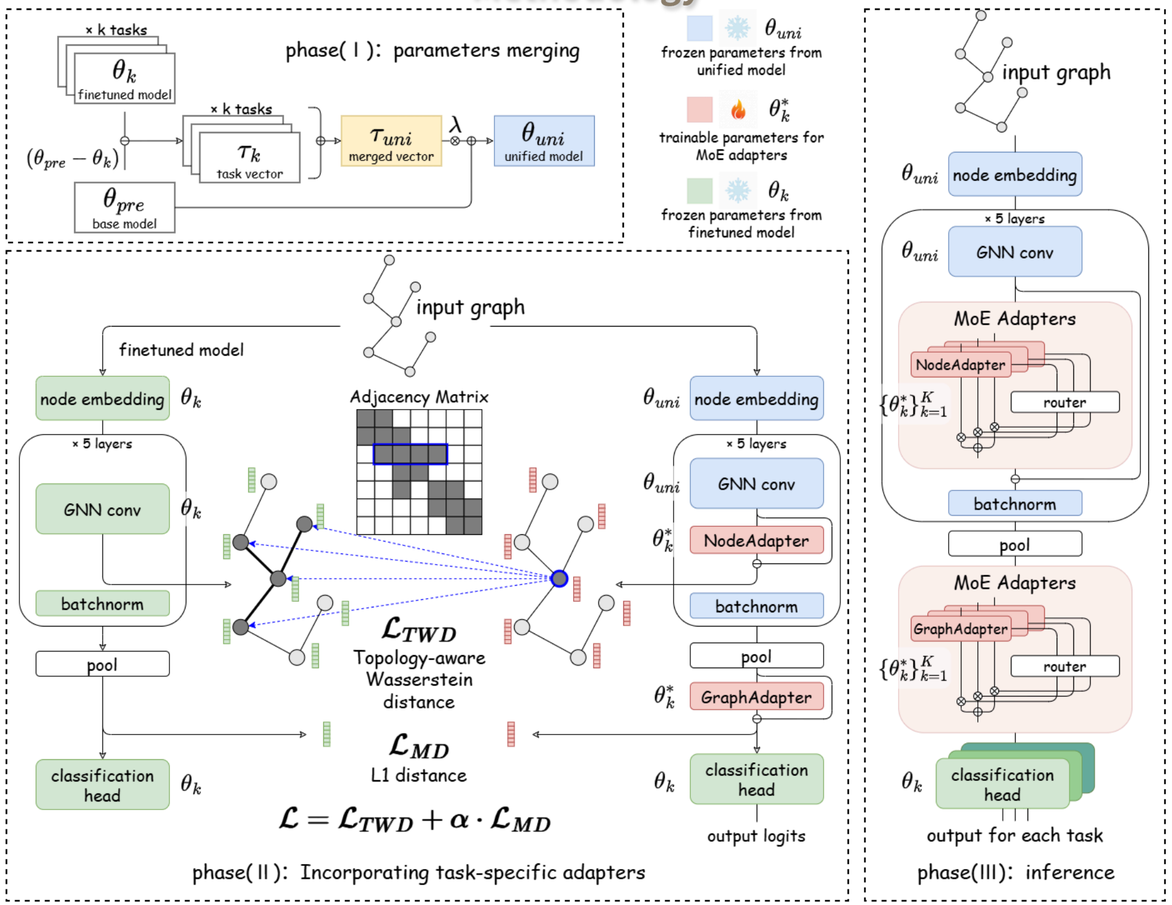


Topological heterogeneity: graph tasks show clearly separated feature distributions due to their distinct topological patterns.

Naive merging fails: simple parameter averaging often performs poorly and loses topology-sensitive knowledge.

Key insight: effective solution should preserve both shared knowledge and task-specific knowledge.

Methodology



1. Parameters Merging: coarsely obtain a unified model by computing the parameters as follows. (shared knowledge)

$$\theta_{uni} = \theta_{pre} + \lambda \sum_{k=1}^{K} (\theta_k - \theta_{pre})$$

2. Training Adapters: two loss functions are used to align the features with those of the fine-tuned model. (task-specific knowledge)

$$L_{TWD} = \mathrm{TWD}(H^{(l)}_{\theta_{uni},\theta_k^*}, H^{(l)}_{\theta_k}, A)$$

$$L_{MD} = \left\| h_{\theta_{uni},\theta_k^*} - h_{\theta_k} \right\|_1$$

$$\min_{\theta_k^*} \frac{1}{|D_k|} \sum_{G \in D_k} (\alpha L_{MD} + \sum_{l=1}^{L} L_{TWD})$$

3. MoE Adapters: training-free router dynamically assigns expert weights based on TWD at inference.

$$\{w_i^{(l)}\}_{i=1}^K = \mathrm{softmax}(\{-\mathrm{TWD}(H_i^{(l)}, H_k^{(l)})\}_{i=1}^K)$$

Experiments

Performance: when merging GCNs and GINs across 8 molecular graph tasks, G-Merging achieves superior multi-task performance.

Router Analysis: router adaptively allocates larger weights to experts from related tasks, enabling cross-task knowledge sharing.

Efficiency: adapters use only 144K parameters, about 7.75% of a full GNN, and training costs roughly 1/8 of full fine-tuning.

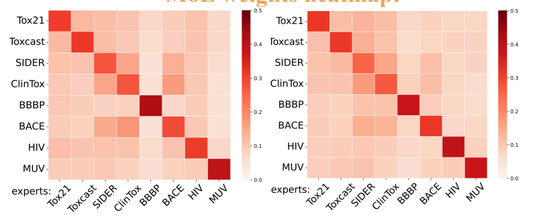
Methods	Tox21	Toxcast	SIDER	ClinTox	BBBP	BACE	HIV	MUV	Average
Full Fine-Tuned	75.8	64.7	60.2	65.2	71.2	76.7	77.0	81.0	71.5
Pretrained	70.4	58.5	56.9	45.4	61.7	70.8	54.7	70.0	61.1
Multi-Task Learning	73.1	62.9	62.0	61.4	68.6	76.1	74.5	73.7	69.0
Weight Average	71.5	63.0	59.8	46.3	66.3	68.5	62.9	71.9	63.8
Task Arithmetic	71.7	63.0	59.9	47.1	66.2	69.2	62.4	72.1	63.9
Ties-Merging	70.4	58.7	57.9	42.6	61.1	72.3	57.8	72.7	61.7
EMR-Merging	**73.8**	61.3	60.8	53.2	**70.3**	73.2	72.7	66.1	66.4
AdaMerging	68.4	59.1	56.3	34.4	61.2	63.7	61.6	65.1	58.7
Twin-Merging	71.4	59.8	58.5	53.3	62.4	57.4	57.4	72.0	61.5
G-Merging-s (Ours)	73.0±0.2	63.1±0.1	61.9*±0.3	62.2±2.0	69.8±0.2	**73.4**±0.3	68.5±1.5	**78.8**±0.5	68.8
G-Merging (Ours)	73.0±0.2	**63.2**±0.1	**62.0***±0.3	**62.6**±4.2	69.8±0.2	73.3±0.6	**68.6**±1.5	**78.8**±0.5	**68.9**

Methods	Tox21	Toxcast	SIDER	ClinTox	BBBP	BACE	HIV	MUV	Average
Full Fine-Tuned	76.1	66.1	64.8	70.0	70.5	86.1	77.6	79.7	73.9
Pretrained	71.6	64.9	60.0	61.6	54.6	76.4	64.4	65.5	64.9
Multi-Task Learning	73.8	64.0	63.3	71.2	68.8	81.1	74.5	72.8	71.2
Weight Average	74.1	**66.1**	62.2	65.9	63.6	78.7	67.0	68.5	68.3
Task Arithmetic	74.1	65.8	63.1	71.6	66.1	75.7	68.9	66.8	69.0
Ties-Merging	71.7	65.0	59.7	61.5	54.8	78.3	64.4	66.6	65.2
EMR-Merging	76.6	65.3	64.1	67.8	70.4	82.1	70.2	67.1	70.4
AdaMerging	71.3	63.6	59.7	72.7	57.3	72.4	68.6	60.8	65.8
Twin-Merging	71.8	64.6	58.8	67.6	56.4	56.7	61.6	69.0	63.3
G-Merging-s (Ours)	**77.3***±0.4	66.0±0.1	64.6±0.2	**74.1***±0.5	**69.2**±0.3	83.1±0.6	**74.9**±0.7	75.5±0.5	**73.1**
G-Merging (Ours)	76.9*±0.4	66.0±0.1	**64.8**±0.2	71.8*±0.4	68.9±0.3	**84.6**±0.5	74.2±0.7	**77.4**±0.6	**73.1**

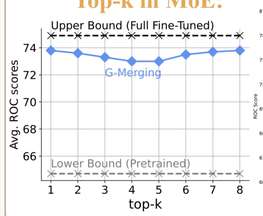
Efficiency:

The total number of parameters	
one full GNN model	1857900
our MoE adapters	144000
running time on single 4090 GPU	
Full finetuning	400+ min
Multi-Task Learning	144 min 45 s
G-Merging (ours)	57 min 56 s

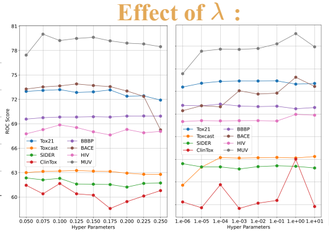
MoE weights heatmap:



Top-k in MoE:



Effect of λ :



Ablations: