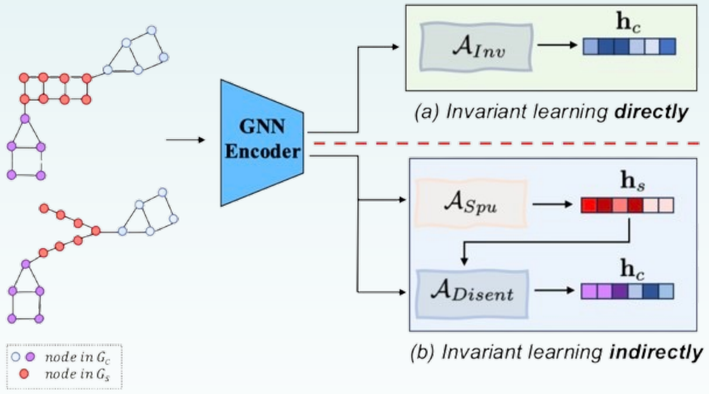
Learning Graph Invariance by Harnessing Spuriosity

Tianjun Yao[1], Yongqiang Chen[1,2], Kai Hu[2], Tongliang Liu[3,1], Kun Zhang[1,2], Zhiqiang Shen[1]

Mohamed bin Zayed University of Artificial Intelligence, Carnegie Mellon University, The University of Sydney

Introduction

- **In real-world applications, it is often the case that multiple invariant features are causally related to the target labels. However, it remains under-explored to what extent traditional (graph) OOD methods are capable of learning these invariant features.**
- **In this work, we find that traditional (graph) OOD methods may only learn a subset of invariant features, hindering their efficacy on OOD generalization performance.**
- **In light of this, we propose LIRS, a learning framework by first learning spurious features, followed by learning invariant features via harnessing the spurious ones.**



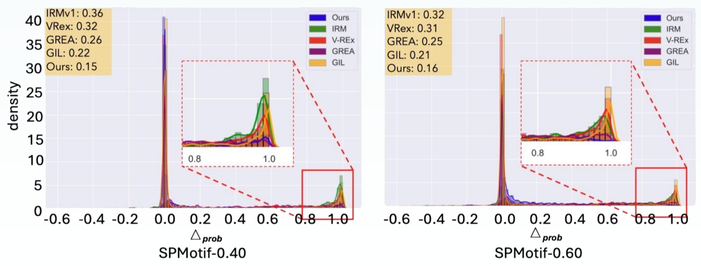


Figure 2: Comparison of the distribution $\mathbb{P}_{\Delta_{prob}}$ across different OOD algorithms.

The Biased Infomax Principle

Definition 2. (*The Infomax Principle*) The infomax principle optimizes the following optimization objective in Eqn. 1 w.r.t the GNN encoder $h_\theta(\cdot)$.

$$\max_{\theta} \mathbb{E}_{G\sim\mathcal{G}} \frac{1}{|G|} \sum_{v_i \in G} I\left(\widehat{\mathbf{h}}_i, \widehat{\mathbf{h}}_G\right), \ s.t.\ \widehat{\mathbf{h}}_i = h_\theta(G), \widehat{\mathbf{h}}_G = READOUT(\widehat{\mathbf{h}}_i). \tag{1}$$

Definition 3. (*The Biased Infomax Principle*) The biased infomax principle optimizes the following optimization objective in Eqn. 2 w.r.t the GNN encoder $h_\theta(\cdot)$.

$$\max_{\theta} \mathbb{E}_{G\sim\mathcal{G}} \frac{1}{|G|} \left(\sum_{v_i \in G_s} I\left(\widehat{\mathbf{h}}_i, \widehat{\mathbf{h}}_G\right) - \sum_{v_i \in G_c} I\left(\widehat{\mathbf{h}}_i, \widehat{\mathbf{h}}_G\right) \right), \text{ s.t. } \widehat{\mathbf{h}}_i = h_\theta(G),\ \widehat{\mathbf{h}}_G = \text{READOUT}(\widehat{\mathbf{h}}_i)$$

- **Biased infomax principle is refined version of Infomax, which adopt contrastive learning for learning spurious features. (see more details in the paper)**
- **Theoretically, we show that biased infomax is the spuriosity learner (Theorem 4.1), which learns spurious features effectively.**

Learning invariant features with Intra-class cross-entropy loss

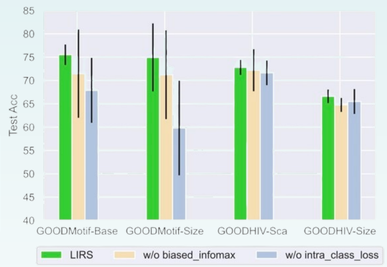
$$\mathcal{L}_{\text{inv}} := \mathbb{E}_{y\sim\mathcal{Y}}\, \mathbb{E}_{G^{(i)}|Y=y} \left(-\sum_{j=1}^{K} w^{(i)} \mathbf{s}_j^{(i)} \log\left(\sigma(\widehat{\mathbf{s}}_j^{(i)})\right), \ s.t., \widehat{\mathbf{s}}^{(i)} = \rho'(\widehat{\mathbf{h}}_G^{(i)}) \in \mathbb{R}^K \right.$$

- **Previous study utilizes cross-entropy loss for feature disentanglement, However, we can show that the vanilla CE loss will fail to identify invariant features when the spurious correlations remain similarly across different classes. To mitigate this weakness, we propose intra-class CE loss.**
- **In Theorem 4.2, we show that there exists a stable solution for $\mathcal{L} = \mathcal{L}_{GT} + \lambda\mathcal{L}_{inv}$ which encodes invariant features.**

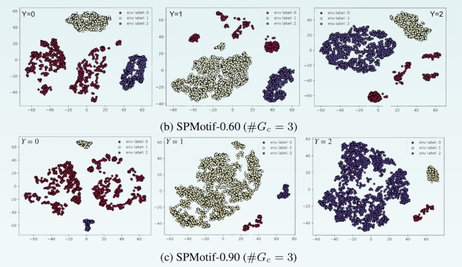
Experiments

Table 1: Performance on synthetic and real-world datasets. Numbers in **bold** indicate the best performance, while the underlined numbers indicate the second best performance.

Method	GOOD-Motif		GOOD-HIV		OGBG-Molbace		OGBG-Molbbbp	
	base	size	scaffold	size	scaffold	size	scaffold	size
ERM	68.66±4.25	51.74±2.88	69.58±2.51	59.94±2.37	75.11±3.03	83.60±3.47	68.10±1.68	78.29±3.76
IRM	70.65±4.17	51.41±3.78	67.97±1.84	59.00±2.92	75.47±2.22	83.12±2.58	67.22±1.15	77.56±2.48
GroupDRO	68.24±8.92	51.95±5.86	70.64±2.57	58.98±2.16		-	66.47±2.39	79.27±2.43
VREx	71.47±6.69	52.67±5.54	70.77±2.84	58.53±2.88	72.81±4.29	82.55±2.51	68.74±1.03	78.76±2.37
RSC	46.12±3.76	51.70±5.47	69.16±3.23	61.17±0.74	74.59±3.65	84.34±2.65	69.01±2.84	78.07±3.89
DiverseModel	54.24±8.22	41.01±1.98	69.17±3.62	61.59±2.23	73.48±3.56	79.40±1.70	68.04±3.27	77.62±1.90
DropEdge	45.08±4.46	45.63±4.61	70.78±1.38	58.53±1.26	70.81±2.12	76.39±2.29	66.49±1.55	78.32±3.44
FLAG	61.12±5.39	51.66±4.14	68.45±2.30	60.59±2.95	80.37±1.58	84.72±0.88	67.69±2.36	79.26±2.26
LiSA	54.59±4.81	53.46±3.41	70.38±1.45	52.36±3.73	78.05±5.01	83.92±2.52	68.11±0.52	78.62±3.74
DIR	62.07±8.75	52.27±4.56	68.07±2.29	58.08±2.31	75.49±2.80	77.42±7.43	66.86±2.25	76.40±4.43
DisC	51.08±3.08	50.39±1.15	68.07±1.75	58.76±0.91	57.78±3.60	71.13±8.86	67.12±2.11	56.59±10.09
CAL	65.63±4.29	51.18±5.60	67.37±3.61	57.95±2.24	76.29±1.60	79.68±4.06	68.06±2.60	79.50±4.81
GREA	56.74±9.23	54.13±10.02	67.79±2.56	60.71±2.20	77.16±1.37	83.15±9.07	69.72±1.66	77.34±3.52
GSAT	62.80±11.41	53.20±8.35	68.66±1.35	58.06±1.98	72.32±5.66	82.45±2.73	66.78±1.45	75.63±3.83
CIGA	66.43±11.31	49.14±8.34	69.40±2.39	59.55±2.56	76.44±1.72	83.95±2.75	64.92±2.09	65.98±3.31
AIA	73.64±5.15	55.85±7.98	71.15±1.81	61.64±3.37	79.42±2.01	85.11±0.74	70.79±1.53	81.03±5.15
OOD-GCL	56.46±4.61	60.23±8.49	70.85±2.07	58.48±2.94	75.96±2.21	85.34±1.77	67.28±3.09	78.11±3.32
EQuAD	67.11±10.11	59.72±3.69	72.24±0.64	64.19±0.56	79.15±2.32	86.41±5.63	70.22±2.36	80.82±5.28
LIRS	**75.51**±2.19	**74.95**±7.69	**72.82**±1.61	**66.64**±1.44	**81.91**±1.98	**88.77**±1.64	**71.04**±0.76	**82.19**±1.57



Ablation study



Visualization of spurious embeddings

- **LIRS achieve superior performance, even in more challenging datasets such as Motif-size.**
- **Intra-class CE loss is more crucial than biased Infomax learning.**
- **Biased Infomax effectively learn spurious features.**

paper

code