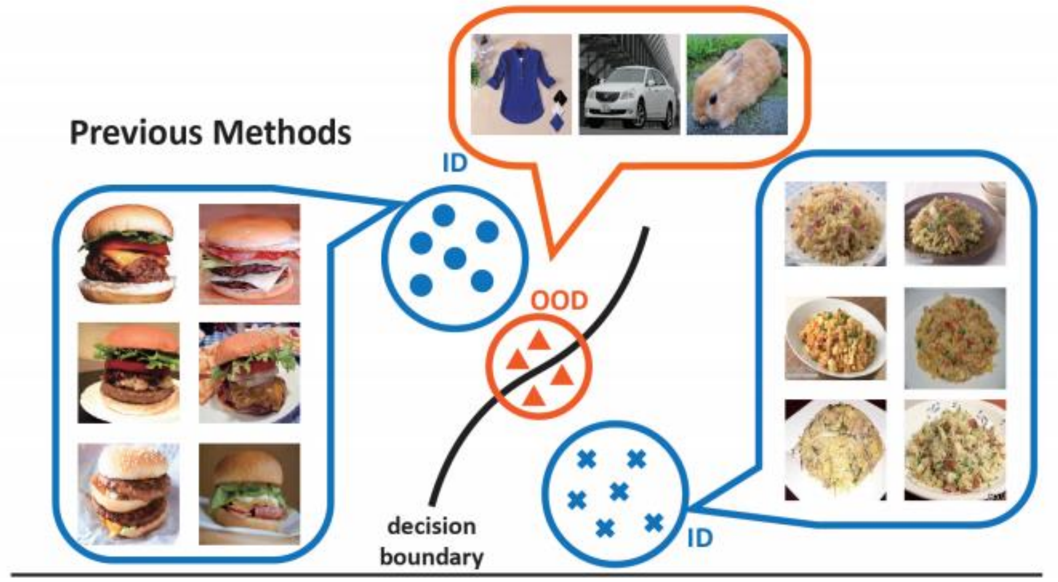
End-to-End Open-Set Semi-Supervised Node Classification with Out-of-Distribution Detection

Tiancheng Huang[1,2,3], Donglin Wang[2,3*], Yuan Fang[4], and Zhengyu Chen[2,3]
1 Zhejiang University, Hangzhou, China
2 Westlake University, Hangzhou, China
3 Westlake Institute for Advanced Study, Hangzhou, China
4 Singapore Management University, Singapore

{huangtiancheng, wangdonglin, chenzhengyu}@westlake.edu.cn, yfang@smu.edu.sg

Motivation

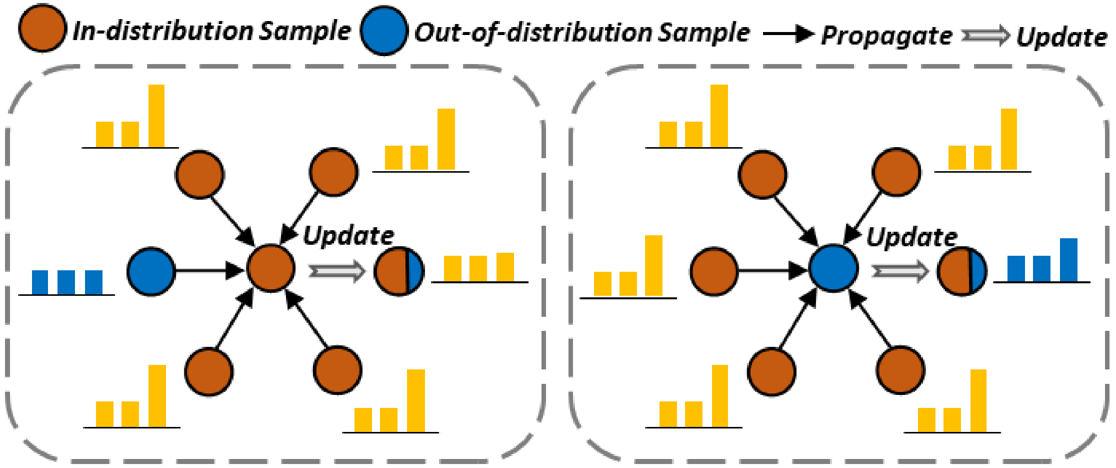
OOD Detection on Images [1]



W/o Propagation and Aggregation between ID and OOD Samples

OOD Detection on Graphs (Ours)

Across-distribution Mixture



(a) Mix to in-distribution (b) Mix to out-of-distribution

W/ Propagation and Aggregation between ID and OOD Samples

[1] Yu Q, Aizawa K. Unsupervised Out-of-Distribution Detection by Maximum Classifier Discrepancy. ICCV 2019

Across-distribution Mixture

Theorem. ***Across-distribution Mixture*** [1]. It is assumed that the sample feature conforms to the Normal distribution with the mean μ and variance σ. The mixing feature of a sample comes from in-distribution and out-of-distribution:

$$\begin{aligned} P_{mix}(\mathbf{x}) &= P(\mathbf{x}|O=in)P(O=in) + P(\mathbf{x}|O=out)P(O=out) \\ &\sim \mathcal{N}(\mu_1, \sigma_1) + \mathcal{N}(\mu_2, \sigma_2), \end{aligned}$$

where $P(\mathbf{x}|O{=}in),\ P(\mathbf{x}|O{=}out) \sim \mathcal{N}(\mu_i, \sigma_i),\ i \in \{1, 2\}.$

Lemma. ***Across-distribution Mixture on Graphs.*** Take one-layer aggregation of GNN as an example, the distribution mixture comes from the central node and its neighbors:

$$P_{mix}(\boldsymbol{x}_i) \sim \mathcal{N}(\mu_i, \sigma_i) + \sum_{v=1}^{|\mathbb{N}(u)|} w_{v,u} \mathcal{N}(\mu_{j_v}, \sigma_{j_v})$$

where $i, j_v \in \{1, 2\},\ w_{v,u}$ denotes weights between node u and v, and $\mathbb{N}(u)$ denotes neighbors of node u

[1] Bitterwolf et al. Revisiting ood detection: A simple baseline is surprisingly effective. ICLR 2022 Submitted.

Unified Learning Framework

To avoid the across-distribution mixture

The joint probability distribution of node label Y and latent variable O

$$P_\theta(Y, O|\mathbf{X}, \mathbf{A}) = P_\theta(Y|\mathbf{X}, \mathbf{A}, O)P(O|\mathbf{X}, \mathbf{A}),$$

i) Learning the GNN parameter by maximizing the likelihood

$$\log P_\theta(Y, O|\mathbf{X}, \mathbf{A}) = \log \sum_k P_\theta(Y|\mathbf{X}, \mathbf{A}, O_k) \cdot P(O_k|\mathbf{X}, \mathbf{A}),$$

ii) Inferring the following posterior of the latent variable O as

$$P_\theta(O_k|\mathbf{X}, \mathbf{A}, Y) = \frac{P_\theta(Y|\mathbf{X}, \mathbf{A}, O_k)}{\sum_j P_\theta(Y|\mathbf{X}, \mathbf{A}, O_j)},$$

Challenges

1) involves marginalizing the latent variable O

2) lacks of supervision for test nodes for inference

Unified Learning Framework

Variational Inference

Introducing variational distribution Q

$$\mathcal{L}(\theta) = \mathbb{E}_{Q(O_k)}\left[\log P_\theta(Y|\mathbf{X}, \mathbf{A}, O_k)\right] - \mathrm{KL}(Q(O_k)||P(O_k)),$$

Introducing the parameterized posterior Q_ϕ with parameter ϕ, and minimizing Kullback-Leibler (KL) divergence, to make the variational distribution Q_ϕ close to its intractable true posterior distribution

$$\mathrm{KL}(Q_\phi\|P) = \mathbb{E}_Q\left[\log \frac{Q_\phi(O_k|\mathbf{X}, \mathbf{A})}{P(O_k)}\right].$$

where P follows Bernoulli distribution

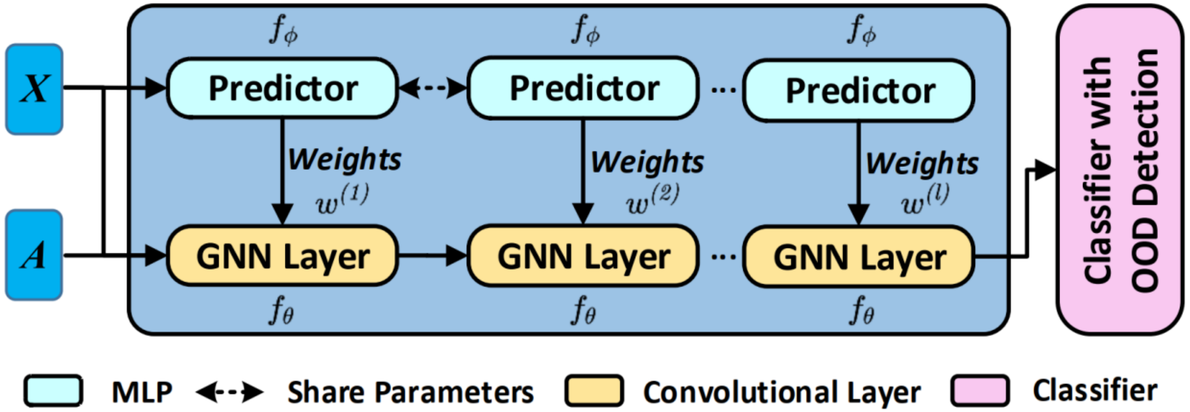
Negative ELBO $\mathcal{L}(\theta, \phi) = -\mathbb{E}_{Q_\phi}[\log P_\theta(Y|\mathbf{X}, \mathbf{A}, O)] + \mathrm{KL}(Q_\phi\|P),$

Learning to Mix Neighbors

Learning and Aggregating Weights

Predictor f_ϕ gives a single scalar between 0 and 1 (parametrized as a sigmoid):

$$Q_\phi(w_{v,u}|\mathbf{X},\mathbf{A}) = \frac{1}{1+exp(-\mathbf{W}^T[\mathbf{H}_v \| \mathbf{H}_u])},$$



Bi-level Optimization

Updating outer level $\quad \min_\phi \mathcal{L}_{val}(\theta^*(\phi), \phi),$

Updating inner level $\quad s.t.\ \theta^*(\phi) = \arg\min_\theta \mathcal{L}_{train}(\theta, \phi),$

Experiments

Datasets. 1) Cora; 2) Citeseer; 3) Pubmed; and 4) ogbn-arXiv

For the split of OOD classes, we strictly follow the standard OOD detection benchmark on graphs [Stadler et al., 2021]. The statistics of datasets are presented in the Table below.

	Cora	**Citeseer**	**Pubmed**	**arXiv**
# Nodes	2,708	3,327	19,717	169,343
# Edges	10,556	9,104	88,648	2,315,598
# Features	1,433	3,703	500	128
# Labels	7	6	3	40
# $\vert\mathcal{C}_{out}\vert$	3	2	1	15
# Fraction	33.38%	33.18%	39.94%	39.11%

Baselines.

1) GCN [Kipf and Welling, 2017], 2) ChebNet [Defferrard et al., 2016], 3) GraphSAGE [Hamilton et al., 2017], 4) GAT [Veliˇckoví c et al., 2018], 5) SGC [Wu et al., 2019], 6) JKNet [Xu et al., 2018], 7) APPNP [Klicpera et al., 2018], 8) SuperGAT [Kim and Oh, 2020], 9) GCNII [Chen et al., 2020], and 10) DropEdge [Rong et al., 2019].

Experiments

Comparison of **semi-supervised node classification** accuracy (%)

Methods	Cora	Citeseer	Pubmed	arXiv
GCN	87.4±0.3	66.0±0.6	89.0±0.2	47.4±0.6
ChebNet	85.6±0.4	65.0±0.6	88.4±0.3	46.5±0.4
GraphSAGE	85.3±1.2	65.8±0.7	89.6±0.6	46.8±0.9
GAT	88.7±0.6	69.6±0.6	90.6±0.9	49.8±1.5
SGC	87.2±0.3	69.2±0.2	91.5±0.6	40.5±2.6
JKNet	86.7±1.1	67.3±0.7	93.1±0.1	50.6±0.6
APPNP	88.2±0.4	68.3±0.5	93.2±0.1	51.3±0.9
SuperGAT	88.3±0.5	69.3±0.8	91.3±1.0	49.2±0.7
GCNII	88.7±0.3	69.4±1.4	93.0±0.7	51.6±1.7
DropEdge	88.9±0.7	69.6±1.2	93.0±0.9	51.7±2.7
LMN(Ours)	**89.7**±0.6	**71.1**±0.6	**93.4**±0.1	**54.1**±1.4

Comparison of **OOD detection** AUROC (%)

Methods	Cora	Citeseer	Pubmed	arXiv
GCN	77.8±0.4	73.1±2.2	63.3±1.4	56.1±0.5
ChebNet	73.5±1.3	69.7±4.0	62.2±1.2	57.1±0.8
GraphSAGE	75.6±1.8	72.8±3.1	59.5±2.0	56.9±1.0
GAT	80.2±1.4	77.9±3.1	61.6±4.2	58.0±1.0
SGC	70.0±0.8	75.5±2.3	61.4±1.8	51.8±1.5
JKNet	76.3±1.8	70.8±3.4	64.4±1.8	52.9±0.6
APPNP	77.8±0.5	72.3±2.7	64.3±0.8	53.7±0.3
SuperGAT	78.5±1.6	78.1±1.6	63.2±3.9	54.1±0.8
GCNII	78.0±1.3	72.4±2.1	65.2±3.9	56.3±2.2
DropEdge	79.3±0.9	75.2±3.5	63.0±2.1	57.9±0.9
LMN(Ours)	**80.5**±1.2	**78.5**±3.2	**68.7**±1.3	**60.4**±0.3

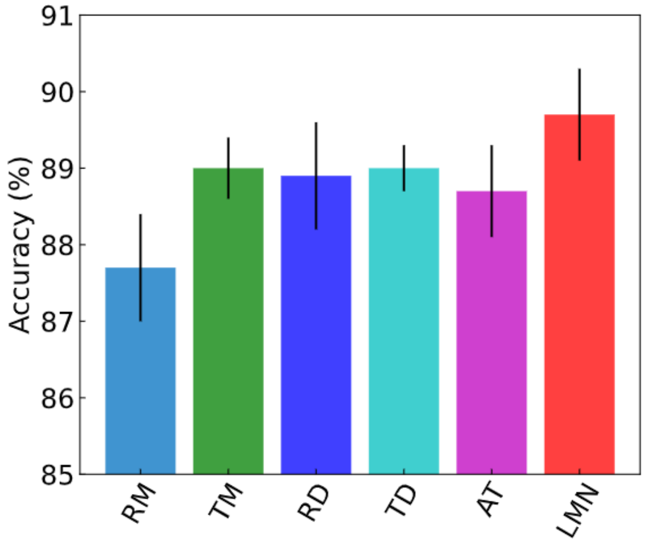
Experiments

1. Mixing Strategies

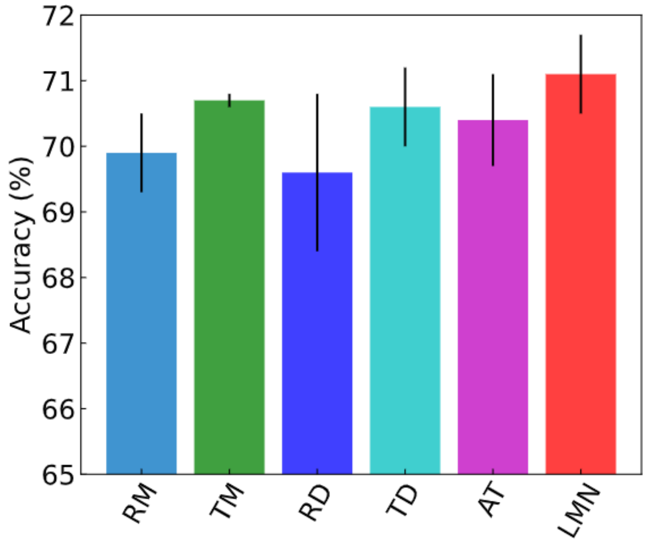
1) RandomMask (RM) 2) TruthMask (TM)

3) RandomDrop (RD) 4) TruthDrop (TD)

5) ATtention (AT) 6) LMN (Ours)



(a) Cora



(b) Citeseer

2. The Effect of Bi-level Optimization

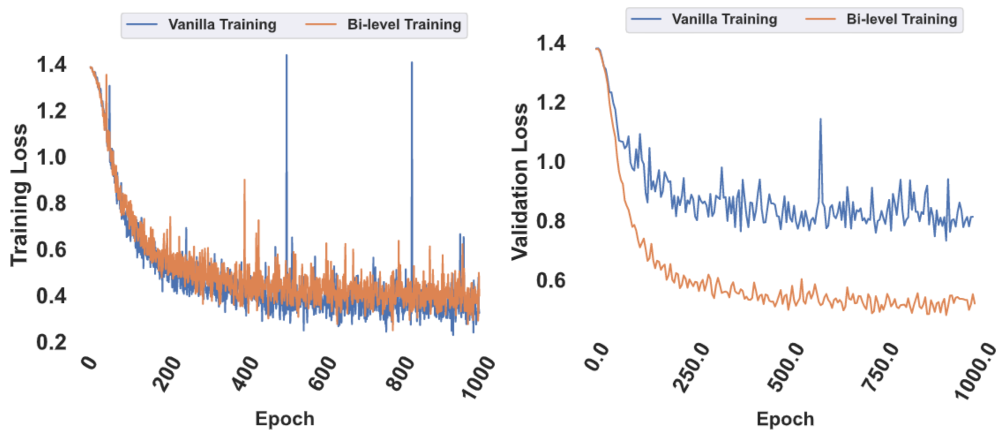


Figure 4: The training and validation losses on Cora.

3. Ablation Study

Methods	OOD Modules	Cora	Citeseer
GCNII	None	$88.7_{\pm 0.3}$	$69.4_{\pm 1.4}$
LMN	**Mixing Neighbors**	$\mathbf{89.7}_{\pm 0.6}$	$\mathbf{71.1}_{\pm 0.6}$

Conclusions

- In this paper, we study a novel problem of end-to-end open-set semi-supervised node classification with OOD detection.
- The novel method LMN in a variational inference framework has been proposed for node classification and OOD detection in an end-to-end manner.
- Extensive experiments on four datasets demonstrate the effectiveness of our proposed method.