Conditional Distribution Learning on Graph Classification —Supplementary Material

Anonymous submission

Technical Appendices

Theorem 1 *Assume that there are n_g graphs and K negative samples for each graph, where $sim(\mathbf{h}_i, \mathbf{h}_k^w) = 0$ $(1 \leq k \leq K)$. Given two conditional distributions $\mathbf{h}_i^w$ and $\mathbf{h}_i^s$ relative to $\mathbf{h}_i$, denoted as $p(\mathbf{h}_i^w|\mathbf{h}_i)$ and $p(\mathbf{h}_i^s|\mathbf{h}_i)$, respectively, and the distribution divergence $\mathcal{L}_d$ in Eq. (11), the following inequality holds:*

$$\mathcal{L}_d \geq \log(K+1) - \frac{1}{\tau}$$

if K satisfies the following condition, i.e.,

$$K \geq e^{\frac{1}{\tau}} - 1$$

where τ is a temperature parameter.

Proof of Theorem 1

Proof *Let $\mathbf{h}_i$, $\mathbf{h}_i^w$ and $\mathbf{h}_i^s$ be the ith columns of $\mathbf{H}$, $\mathbf{H}_w$ and $\mathbf{H}_s$, respectively. Given two conditional distributions $p(\mathbf{h}_i^w|\mathbf{h}_i)$ and $p(\mathbf{h}_i^s|\mathbf{h}_i)$, we have*

$$0 < p(\mathbf{h}_i^w|\mathbf{h}_i) \leq 1 \quad \textit{and} \quad 0 < p(\mathbf{h}_i^s|\mathbf{h}_i) \leq 1$$

where $1 \leq i \leq n_g$. Let

$$l = p(\mathbf{h}_i^w|\mathbf{h}_i) \log\left(p(\mathbf{h}_i^s|\mathbf{h}_i)\right)$$

and we obtain

$$l \leq \frac{1}{\tau} - \log(K+1).$$

if

$$K \geq e^{\frac{1}{\tau}} - 1$$

Hence,

$$\mathcal{L}_d \geq \log(K+1) - \frac{1}{\tau}.$$

Analysis of Augmented Views

Definition 1 (Sufficient Augmented View) *Given an encoder f, an augmented view v_{suf} is sufficient in contrastive learning if and only if $I(v; y) = I(f(v_{suf}); y)$, where v represents an original view, and $I(\cdot;\cdot)$ denotes mutual information.*

Intuitively, the augmented view v_{suf} is sufficient for predicting the target label y if all the information in v_{suf} is preserved at approximately v during the graph embedding encoding phase. Thus, $f(v_{suf})$ contains all the shared information between v_{suf} and v. This indicates that $f(v_{suf})$ keeps all the task-relevant information from the original view.

Definition 2 (Minimal Sufficient Augmented View) *Among all sufficient augmented views, a view v_{min} is minimal if and only if*

$$I(f(v_{min}); y) \leq I(f(v_{suf}); y)$$

for all sufficient views v_{suf}.

Theorem 2 *Graph representations obtained by GNNs are employed to predict the target label y in a graph classification task. The minimal sufficient view v_{min} contains less task-relevant information from the original view v than other sufficient view v_{suf}. Thus, we have*

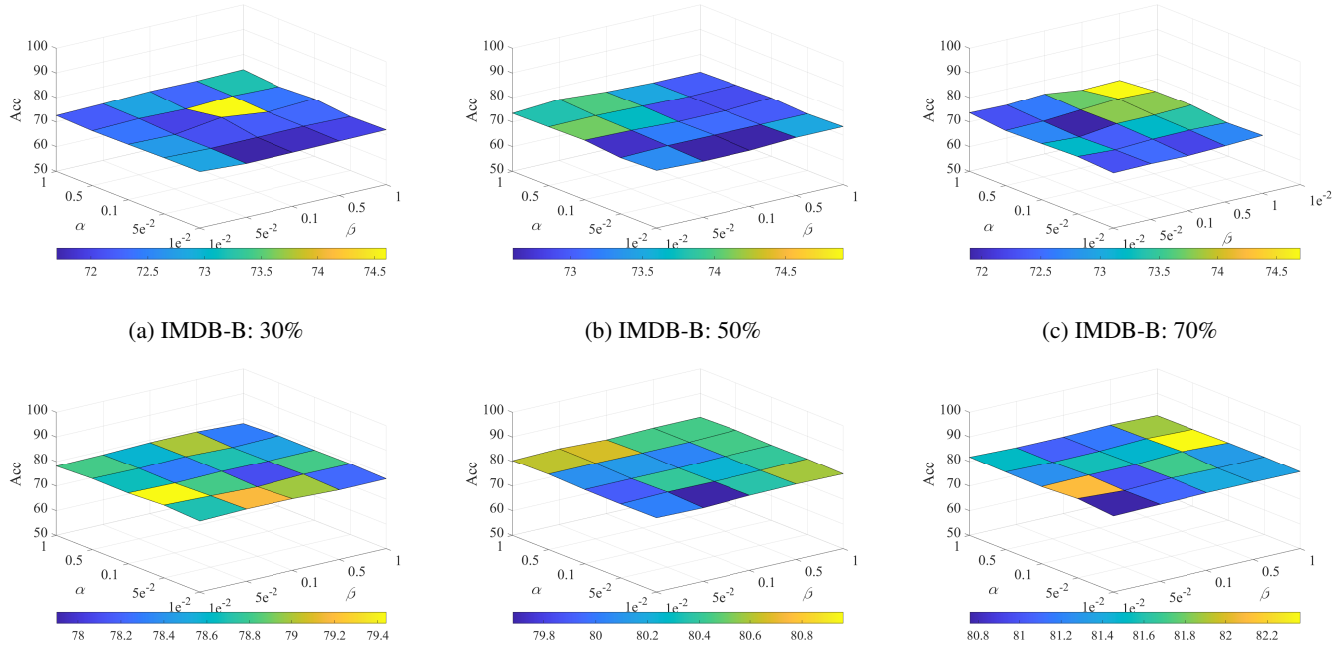
$$I(v_{suf}; y) \geq I(f(v_{min}); y)$$

Proof

$$\begin{aligned}
&I(f(v_{suf}), f(v_{min}); y) \\
=&I(f(v_{min}); y) + I(f(v_{suf}); y|f(v_{min})) \\
\Rightarrow&I(f(v_{suf}); y) = I(f(v_{min}); y) + I(f(v_{suf}); y|f(v_{min})) \\
\Rightarrow&I(v_{suf}; y) = I(f(v_{min}); y) + I(f(v_{suf}); y|f(v_{min})) \\
\Rightarrow&I(v_{suf}; y) \geq I(f(v_{min}); y)
\end{aligned}$$

Among all sufficient augmented views, the minimal sufficient view v_{min} contains the least information about v_{suf}. It is assumed that v_{min} contains only the information shared between v_{suf} and v. This implies that v_{min} eliminates the information that is not shared between v_{suf} and v. However, some task-relevant information might not be present in the shared information between views (Wang et al. 2022; Tian et al. 2020). According to Theorem 2, v_{suf} contains more task-relevant information. CDL is an end-to-end graph representation learning model. Therefore, v_{suf} provides more valuable information than does v_{min} for CDL.

Let v_{min} and v_{suf} be a strongly augmented view and a weakly augmented view, respectively. The view v_{min} introduces diversity to graph-structured data, which can enhance the generalizability of GNNs. In contrast, the view



(a) IMDB-B: 30% (b) IMDB-B: 50% (c) IMDB-B: 70%

(d) IMDB-B: 30% (e) IMDB-B: 50% (f) IMDB-B: 70%

Figure 1: Graph classification results with different α and β combinations across different percentages of training samples.

Table 1: Statistics of the graph datasets.

Datasets	#Class	#Graph	Avg. #Node	Avg. #Edge
MUTAG	2	188	17.93	19.79
PROTEINS	2	1113	39.06	72.82
IMDB-B	2	1,000	19.77	96.53
NCI1	2	4, 110	29.87	32.30
RDT-B	2	2,000	429.63	497.75
RDT-M5K	5	4,999	508.52	594.87
COLLAB	3	5,000	74.49	2,457.2
GITHUB	2	12,725	113.79	236.64

v_{suf} contains nonshared information between v_{min} and v, which may be crucial for graph classification tasks. This demonstrates that the distribution divergence $\mathcal{L}_d$ in Eq. (11) effectively leverages both the diversity and the quantity of data provided by graph-structured data augmentations, while preserving the intrinsic semantic information.

Experimental details

In this section, we evaluate the performance of the proposed CDL method with benchmark datasets. All the experiments are conducted on a Linux workstation with a GeForce RTX 4090 GPU (24 GB caches), an Intel(R) Xeon(R) Platinum 8336C CPU and 128.0 GB of RAM.

Experimental settings

The statistics of the eight benchmark graph datasets are summarized in Table 1. The numbers of classes and graphs in these graph datasets range from 2 to 5 and 188 to 12,725, respectively. In particular, the node masking scheme is applied to both the weak and strong augmentations of the graph-structured data.

Parameter Settings The learning rate for the proposed CDL model was empirically set to $5e^{-3}$. The size of the hidden layers was chosen from the set $\{128, 64\}$. The batch size of the graphs was set to 128 during training and testing, with a dropout of 0.2. The numbers of GNN and MLP layers ranged from 2 to 3 and 1 to 2, respectively. The overall loss of the proposed CDL model in Eq. (14) involves two parameters, α and β, which were selected from $\{0.01, 0.05, 0.1, 0.5, 1\}$ via a grid search strategy. For a fair comparison, the best results of all competing methods are reported after tuning their parameters.

Parameter Sensitivity Analysis

We first perform parameter sensitivity analysis on two critical parameters, α and β, from Eq. (14). These parameters were selected from the set $\{0.01, 0.05, 0.1, 0.5, 1\}$ for CDL. The remaining hyperparameters in the proposed CDL method are determined by the parameter settings. Owing to space limitations, we conducted experiments using two representative datasets, i.e., the IMDB-B and COLLAB datasets. Fig. 1 shows the graph classification results with the IMDB-B and COLLAB datasets with different α and β combinations across different percentages of training samples, respectively. The classification results of the CDL method fluctuate slightly with different α and β combinations. This finding indicates that the proposed CDL method

Table 2: Graph classification results (average accuracy (%) $\pm$ standard deviation (%)) on pairs of strongly augmented views with eight benchmark graph datasets.

Label	Methods	MUTAG	PROTEINS	IMDB-B	NCI1	RDT-B	RDT-M5K	COLLAB	GITHUB
30%	$\text{CDL}_{\mathbf{0.4}}$	87.28±7.48	75.48±5.32	72.30±5.03	78.10±2.21	90.50±1.84	54.07±2.19	78.90±2.01	69.53±1.31
	$\text{CDL}_{\mathbf{0.5}}$	85.12±9.92	75.20±3.14	71.40±4.12	78.00±2.69	90.05±1.46	53.93±2.26	78.50±0.84	69.29±1.27
	$\text{CDL}_{\mathbf{0.6}}$	85.70±7.89	75.48±3.73	70.60±6.55	77.47±2.22	89.90±1.45	54.01±2.31	78.44±1.47	69.35±1.19
	$\text{CDL}_{\mathbf{0.7}}$	85.03±8.77	75.39±3.89	70.40±5.85	77.18±1.69	90.35±2.15	53.89±2.55	78.26±2.07	69.27±1.39
	CDL	**89.36±6.14**	**76.74±4.51**	**74.60 ± 4.25**	**79.37 ±1.62**	**91.15±1.76**	**55.31±1.65**	**79.44±1.82**	**70.15±1.27**
50%	$\text{CDL}_{\mathbf{0.4}}$	88.80±7.09	75.12±4.13	73.40±5.50	78.10±1.82	90.95±1.77	55.09±1.56	80.20±2.12	70.30±0.80
	$\text{CDL}_{\mathbf{0.5}}$	88.27±7.09	74.94±4.66	73.30±5.64	77.20±2.21	91.00±1.89	55.87±2.09	79.94±2.01	70.23±1.27
	$\text{CDL}_{\mathbf{0.6}}$	88.33±8.21	74.49±3.27	72.60±5.82	77.98±1.61	90.70±2.25	55.79±1.56	79.76±2.04	69.60±1.14
	$\text{CDL}_{\mathbf{0.7}}$	88.30±8.31	74.58±3.54	72.30±4.67	75.89±1.34	90.50±2.54	55.17±2.06	78.94±1.50	69.88±2.12
	CDL	**89.94±8.76**	**76.10±2.80**	**74.90±3.70**	**79.08±1.86**	**92.05±2.14**	**56.51±2.32**	**80.96±1.29**	**70.83±1.13**
70%	$\text{CDL}_{\mathbf{0.4}}$	87.81±6.05	75.75±3.79	73.80±5.33	81.14±1.90	91.55±1.38	55.73±2.40	81.36±1.51	70.16±1.47
	$\text{CDL}_{\mathbf{0.5}}$	87.28±8.97	75.39±3.37	73.40±5.64	80.51±1.43	91.60±1.71	54.89±2.24	81.28±1.69	70.15±1.15
	$\text{CDL}_{\mathbf{0.6}}$	86.75±8.36	75.57±3.00	73.30±5.12	80.61±2.06	90.65±2.14	54.39±1.52	81.10±1.70	70.04±0.87
	$\text{CDL}_{\mathbf{0.7}}$	85.70±7.89	74.76±3.12	72.90±5.76	80.66±1.74	90.50±2.15	53.17±2.44	80.82±1.71	69.88±1.48
	CDL	**89.91±7.30**	**77.27±3.62**	**74.90±5.63**	**82.36±1.52**	**92.35±1.63**	**56.65±1.87**	**82.36±1.72**	**71.06±1.17**

usually achieves satisfactory classification results with relatively wide ranges of α and β values.

Evaluation on Pairs of Strongly Augmented Views

The intrinsic semantic information in graph-structured data may be disrupted when strong augmentations introduce significant perturbations. However, strong augmentations substantially enhance the diversity of graph-structured data, which in turn helps improve the generalization ability of GNNs. We further evaluate the performance of the proposed CMD method when applying only strong augmentations to graph-structured data. The masking ratio of the node attributes for the strong augmentation was selected from the set $\{0.4, 0.5, 0.6, 0.7\}$. Specifically, each pair of strongly augmented views on the graph-structured data shares the same masking ratio. The proposed CDL method with masking ratios $\{0.4, 0.5, 0.6, 0.7\}$ are denoted as $\text{CDL}_{\mathbf{0.4}}$, $\text{CDL}_{\mathbf{0.5}}$, $\text{CDL}_{\mathbf{0.6}}$ and $\text{CDL}_{\mathbf{0.7}}$, respectively.

Table 2 shows graph classification results on pairs of strongly augmented views. We observe that the average classification accuracy of the proposed CDL method generally decreases as the masking ratio increases from 0.4 to 0.7. For example, on three different percentages of the dataset set, 30%, 50% and 70%, CDL improves the average classification accuracy by 2.08%, 1.14%, and 2.10% on MUTAG, and by 1.62%, 0.53%, and 0.91% on GITHUB, compared to $\text{CDL}_{\mathbf{0.4}}$. As expected, $\text{CDL}_{\mathbf{0.7}}$ almost achieves the lowest average classification accuracy at the masking ratio of 0.7. The gap in average classification accuracy between CDL and $\text{CDL}_{\mathbf{0.7}}$ has further widened. This demonstrates that strong augmentations tend to disrupt the intrinsic semantic information in graph-structured data. Moreover, the proposed CDL method outperforms all its variants with different masking ratios. This empirically validates that conditional distribution learning enhances the generalizability and robustness of CDL by aligning the conditional distributions of weakly and strongly augmented node embeddings given the original node embeddings.

Discussion

Unlike data augmentation strategies in most existing graph contrastive learning methods, the primary goal of data augmentation in the proposed CDL method is to align the conditional distributions of weakly and strongly augmented features with the original features. The strong augmentations in CDL introduce significant perturbations to the graph-structured data. This enhances the diversity of node embeddings. Conditional distribution learning enhances the ability of the CDL model to capture intrinsic semantic information. As a result, this approach significantly improves the robustness and generalization of the CDL model, while effectively reducing the risk of disrupting intrinsic semantic information. In contrast, data augmentation techniques in other graph contrastive learning methods often lead to ambiguities that may compromise intrinsic semantic information. Although CDL has demonstrated effectiveness in graph classification, there are still several limitations to be solved in the future. For example, the estimation of the node masking ratio for weak augmentation poses a significant challenge. Furthermore, it would be worthwhile to integrate an adaptive augmentation scheme of the graph-structured data with CDL. This potentially yields significant improvements in the stability of computational performance.

References

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning. In *Proceedings of the 36th Advances in Neural Information Processing Systems*, 6827–6839. Vancouver, Canada.

Wang, H.; Guo, X.; Deng, Z.-H.; and Lu, Y. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 16041–16050. New Orleans, Louisiana, USA.