Supplementary Material for A Graph Contrastive Learning Framework with Adaptive Augmentation and Encoding for Unaligned Views

1 G-EMD-based Contrastive Loss

After obtaining the representations of the two unaligned views, we use the g-EMD distance in Rosa as the contrastive metric to measure the distance between positive and negative samples. Most previous contrastive learning studies have used cosine similarity as the contrastive metric. However, cosine similarity is limited to the contrast between aligned views, which inevitably limits the diversity and flexibility of view sampling and augmentation and affects the expressiveness of contrastive learning. Therefore, we use g-EMD distance to solve this problem.

EMD is the measure of the distance between two discrete distributions, it can be interpreted as the minimum cost to move one pile of dirt to the other. Since g-EMD distance can directly calculate the distance between representations of views, it can solve the problem that nodes between views must be aligned. The calculation of g-EMD can be formulated as a linearoptimization problem. In the contrastive of positive and negative samples, the two augmented views have feature mappings $\mathbf{X} \in \mathbb{R}^{M \times d}$ and $\mathbf{Y} \in \mathbb{R}^{N \times d}$, respectively, and the goal is to measure the distance from converting X to Y. Assume that for each node $\boldsymbol{x}_i \in \mathbb{R}^d$, it has $\boldsymbol{t}_i$ units to transport and that node $\boldsymbol{y}_j \in \mathbb{R}^d$ has $\boldsymbol{r}_j$ units to receive. For a given pair of nodes $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$, the unit transport cost is $\mathbf{D}_{ij}$ and the transport volume is $\boldsymbol{\Gamma}_{ij}$. we define the problem as follows:

$$\begin{aligned} &\min_{\mathbf{\Gamma}} \sum_i^M \sum_j^N \mathbf{D}_{ij} \mathbf{\Gamma}_{ij} \\ \text{s.t. } &\mathbf{\Gamma}_{ij} \geq 0, i = 1, 2, \ldots, M, j = 1, 2, \ldots, N \\ &\sum_i^M \mathbf{\Gamma}_{ij} = r_j, j = 1, 2, \ldots, N \\ &\sum_j^N \mathbf{\Gamma}_{ij} = t_i, i = 1, 2, \ldots, M. \end{aligned} \tag{1}$$

We calculate the cost matrix of transferring $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$ by the following equation.

$$\mathbf{D}_{ij} = 1 - \frac{x_i^T y_j}{\parallel x_i \parallel \parallel y_j \parallel} \tag{2}$$

To better fit the graph structure data, we also incorporated the topological distances from the graph in calculating the cost matrix, which formed a modified cost matrix after fusion. We obtained a scaling factor by fusing the activation function with the temperature on the topological distance to the original cost matrix. The scaling factor is calculated as follows. Where s denotes the *sigmoid* nonlinear activation function, Ψ denotes the topological distance, where $\tau \geq 1$is the temperature factor that controls the rate of change of the curve.

$$\mathbf{S}_{i,j} = S\left(\Psi_{i,j}\right) = \frac{1}{1 + e^{-\Psi_{i,j}/\tau}} \tag{3}$$

After obtaining the topological distance-based scaling factor, we fuse it with the original cost matrix to obtain the fused cost matrix. Where $\circ$ is the Hadamard product.

$$\mathbf{D} = 1 - \mathbf{D} \circ \mathbf{S} \tag{4}$$

After obtaining the cost matrix incorporating the topological distance, we have to find the optimal transport $\boldsymbol{\Gamma}$ to compute the g-EMD distance. For this purpose, we can find the optimal transport and simultaneously satisfy the uniform distribution by applying the *Sinkhorn algorithm* with an entropy regularizer.

$$\mathrm{g-EMD}(\mathbf{X}, \mathbf{Y}, \mathbf{S}) = \inf_{\mathbf{\Gamma} \in \Pi} \langle \mathbf{\Gamma}\ , \mathbf{D}_{\mathrm{F}} \rangle + \underbrace{\frac{1}{\lambda} \mathbf{\Gamma}(\log \mathbf{\Gamma} - 1)}_{\text{regularizationterm}} \tag{5}$$

Where $\langle , \rangle_{\mathrm{F}}$ denotes the Frobenius inner product, which is the hyperparameter controlling the strength of the regularization. By regularization, we can get the optimal $\boldsymbol{\Gamma}$ as follows, where α and β are two coefficient vectors.

$$\tilde{\Gamma} {=} \alpha_{\mathrm{i}} \beta_j e^{-\lambda D_{ij}} \tag{6}$$

After obtaining the cost matrix and the optimal cost matrix, we can calculate the g-EMD distance to convert X to Y.

$$\mathrm{g-EMD}(\mathbf{X}, \mathbf{Y}, \mathbf{S}) = \langle \tilde{\boldsymbol{\Gamma}}, \mathbf{D} \rangle_{\mathrm{F}} \tag{7}$$

In order to map the node representations of different views into the same contrastive space, we send the obtained node representations into a projection head (i.e., a two-layer MLP) to obtain $\mathrm{Z}_1^{(n)}$, $\mathrm{Z}_2^{(n)}$. Our loss function is as follows. Where where s(x,y)=g-EMD(x,y) is used to calculate the similarity between x and y, $\mathbb{I}$ is an indicator function that returns 1 if $i = k$; otherwise returns 0, τ is the temperature parameter.

$$\ell\left(\mathbf{Z}_1^{(i)}, \mathbf{Z}_2^{(i)}\right) = -\log\left(\frac{e^{s\left(\mathbf{z}_1^{(i)}, \mathbf{z}_2^{(i)}\right)\big)/\tau}}{\sum\limits_{k=1}^{N} e^{\mathrm{s}\left(\mathbf{z}_1^{(i)}, \mathbf{z}_2^{(k)}\right)\big)/\tau} + \sum\limits_{k=1}^{N} \mathbb{I}_{[k \neq i]} e^{\mathrm{s}\left(\mathbf{z}_1^{(i)}, \mathbf{z}_1^{(k)}\right)\big)/\tau}}\right), \tag{8}$$

2 Datasets.

We conducted experiments on seven public benchmark datasets, including four homophilic datasets, Cora, Citseer, Amazon-Photo, and Amazon-Computers, and three heterophilic datasets, Cornell, Texas, Wisconsin. We tested the performance of the xxxx model on the node classification task. All the datasets we used are from the Pytorch Geometry Library (PyG). For more information about the above datasets in Table 1.

Table 1. Statistics of datasets used in experiments.

Dataset	#Nodes	#Edges	#Features	#Classes
Cora	2,708	5,429	1,433	7
Citeseer	3,327	4,732	3703	6
Amazon-Photo	7,650	119,081	745	8
Amazon-Computers	13,752	245,861	767	10
Cornell	183	280	1,703	5
Texas	183	295	1,703	5
Wisconsin	251	466	1,703	5

- Cora and Citeseer are two citation network datasets where the nodes in the graph represent various papers, and the edges indicate the citation relationships between papers. The node features are represented as bag-of-words models of the corresponding papers, and the labels are the academic topics of the papers.
- Amazon-Photo and Amazon-Computers are two co-purchasing relationship network datasets from Amazon. The nodes in the graph represent products, and the edges indicate the relationship between two products that are often purchased together. Product reviews are coded as node features and labels are product categories.
- Cornell, Texas, and Wisconsin are the three heterophilic datasets collected by the CMU webKB project. These three datasets are web data collected from the scientific departments of the respective universities, where nodes represent web pages and edges represent hyperlinks between web pages. Node features are represented as bags of words for web pages, and these nodes are manually classified into five categories: students, teachers, programs, courses, and employees.

3 Sensitivity Analysis

In this section, we perform sensitivity analysis on the model by varying the hyperparameters controlling the edge dropping probability p_e and the hyperparameters controlling the feature masking dimension p_f. For comparison, we set $p_e = p_{e,1} = p_{e,2}$ and $p_f = p_{f,1} = p_{f,2}$ to experiment on the Cora dataset. The experimental results are shown in Figure 1. From the experimental results, we can see that if the model does not remove a large number of edges and mask a large number of features dimensions during subgraph augmentation, the hyperparameters do not have a significant impact on the performance of the model, which can indicate that our model is not particularly sensitive to hyperparameters. Suppose larger hyperparameters are set for subgraph augmentation. In that case, important knowledge in the graph may be deleted together, which is not conducive to the model of extracting the semantics of the graph and learning the important knowledge in the graph, so the performance will be degraded more.

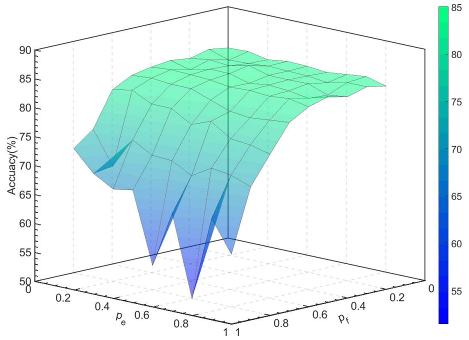

Fig. 1. Node classification accuracies of GCAUV on Cora with different p_e and p_f