ITERATIVE GRAPH SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

How to discriminatively vectorize graphs is a fundamental challenge that attracts increasing attentions in recent years. Inspired by the recent success of unsupervised contrastive learning, we aim to learn graph-level representation in an unsupervised manner. Specifically, we propose a novel unsupervised graph learning paradigm called Iterative Graph Self-Distillation (IGSD) which iteratively performs the teacher-student distillation with graph augmentations. Different from conventional knowledge distillation, IGSD constructs the teacher with an exponential moving average of the student model and distills the knowledge of itself. The intuition behind IGSD is to predict the teacher network representation of the graph pairs under different augmented views. As a natural extension, we also apply IGSD to semi-supervised scenarios by jointly regularizing the network with both supervised and unsupervised contrastive loss. Finally, we show that finetuning the IGSD-trained models with self-training can further improve the graph representation power. Empirically, we achieve significant and consistent performance gain on various graph datasets in both unsupervised and semi-supervised settings, which well validates the superiority of IGSD.

1 INTRODUCTION

Graphs are ubiquitous representations encoding relational structures across various domains. Learning low-dimensional vector representations of graphs is critical in various domains ranging from social science (Newman & Girvan, 2004) to bioinformatics (Duvenaud et al., 2015; Zhou et al., 2020). Many graph neural networks (GNNs) (Gilmer et al., 2017; Kipf & Welling, 2016; Xu et al., 2018) have been proposed to learn node and graph representations by aggregating information from every node's neighbors via non-linear transformation and aggregation functions. However, the key limitation of existing GNN architectures is that they often require a huge amount of labeled data to be competitive but annotating graphs like drug-target interaction networks is challenging since it needs domain-specific expertise. Therefore, unsupervised learning on graphs has been long studied, such as graph kernels (Shervashidze et al., 2011) and matrix-factorization approaches (Belkin & Niyogi, 2002).

Inspired by the recent success of unsupervised representation learning in various domains like images (Chen et al., 2020a; He et al., 2020) and texts (Radford et al., 2018), most related works in the graph domain either follow the pipeline of unsupervised pretraining (followed by fine-tuning) or *InfoMax principle* (Hjelm et al., 2018). The former often needs meticulous designs of pretext tasks (Hu et al., 2019; You et al., 2020) while the latter is dominant in unsupervised graph representation learning, which trains encoders to maximize the mutual information (MI) between the representations of the global graph and local patches (such as subgraphs) (Veličković et al., 2018; Sun et al., 2019; Hassani & Khasahmadi, 2020). However, MI-based approaches usually need to sample subgraphs as local views to contrast with global graphs. And they usually require an additional discriminator for scoring local-global pairs and negative samples, which is computationally prohibitive (Tschannen et al., 2019). Besides, the performance is also very sensitive to the choice of encoders and MI estimators (Tschannen et al., 2019). Moreover, MI-based approaches cannot be handily extended to the semi-supervised setting since local subgraphs lack labels that can be utilized for training. Therefore, we are seeking an approach that learns the entire graph representation by contrasting the whole graph directly without the need of MI estimation, discriminator and subgraph sampling.

Motivated by recent progress on contrastive learning, we propose the Iterative Graph Self-Distillation (IGSD), a teacher-student framework to learn graph representations by contrasting graph instances directly. The high-level idea of IGSD is based on graph contrastive learning where we pull sim-

ilar graphs together and push dissimilar graph away. However, the performance of conventional contrastive learning largely depends on how negative samples are selected. To learn discriminative representations and avoid collapsing to trivial solutions, a large set of negative samples (He et al., 2020; Chen et al., 2020a) or a special mining strategy (Schroff et al., 2015) are necessary. In order to alleviate the dependency on negative samples mining and still be able to learn discriminative graph representations, we propose to use self-distillation as a strong regularization to guide the graph representation learning.

In the IGSD framework, graph instances are augmented as several views to be encoded and projected into a latent space where we define a similarity metric for consistency-based training. The parameters of the teacher network are iteratively updated as an exponential moving average of the student network parameters, allowing the knowledge transfer between them. As merely small amount of labeled data is often available in many real-world applications, we further extend our model to the semi-supervised setting such that we can effectively utilize graph-level labels while considering arbitrary amounts of positive pairs belonging to the same class. Moreover, in order to leverage the information from pseudo-labels with high confidence, we develop a self-training algorithm based on the supervised contrastive loss to fine-tune the encoder.

We experiment with real-world datasets in various scales and compare the performance of IGSD with state-of-the-art graph representation learning methods. Experimental results show that IGSD achieves competitive performance in both unsupervised and semi-supervised settings with different encoders and data augmentation choices. With the help of self-training, our performance can exceed state-of-the-art baselines by a large margin.

To summarize, we make the following contributions in this paper:

- We propose a self-distillation framework called IGSD for unsupervised graph-level representation learning where the teacher-student distillation is performed for contrasting graph pairs under different augmented views.
- We further extend IGSD to the semi-supervised scenario, where the labeled data are utilized effectively with the supervised contrastive loss and self-training.
- We empirically show that IGSD surpasses state-of-the-art methods in semi-supervised graph classification and molecular property prediction tasks and achieves performance competitive with state-of-the-art approaches in unsupervised graph classification tasks.

2 PRELIMINARIES

2.1 FORMULATION

Unsupervised Graph Representation Learning Given a set of unlabeled graphs $\mathcal{G} = \{G_i\}_{i=1}^N$, we aim at learning the low-dimensional representation of every graph $G_i \in \mathcal{G}$ favorable for downstream tasks like graph classification.

Semi-supervised Graph Representation Learning Consider a whole dataset $\mathcal{G} = \mathcal{G}_L \cup \mathcal{G}_U$ composed by labeled data $\mathcal{G}_L = \{(G_i, y_i)\}_{i=1}^l$ and unlabeled data $\mathcal{G}_U = \{G_i\}_{i=l+1}^{l+u}$ (usually $u \gg l$), our goal is to learn a model that can make predictions on graph labels for unseen graphs. And with K augmentations, we get $\tilde{\mathcal{G}}_L = \{\tilde{G}_k\}_{k=1}^{Kl}$ and $\tilde{\mathcal{G}}_U = \{(\tilde{G}_k, \tilde{y}_k)\}_{k=l+1}^{K(l+u)}$ as our training data.

2.2 GRAPH REPRESENTATION LEARNING

We represent a graph instance as $G(\mathcal{V}, \mathcal{E})$ with the node set $\mathcal{V}$ and the edge set $\mathcal{E}$. The dominant ways of graph representation learning are graph neural networks with neural message passing mechanisms (Hamilton et al., 2017): for every node $v \in \mathcal{V}$, node representation $\mathbf{h}_v^k$ is iteratively computed from the features of their neighbor nodes $\mathcal{N}(v)$ using a differentiable aggregation function. Specifically, at the iteration k we get the node embedding as:

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}^k \cdot \mathrm{CONCAT}\left(\mathbf{h}_v^{k-1}, \mathrm{AGGREGATE}_k\left(\left\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\right\}\right)\right)\right) \tag{1}$$

Then the graph-level representations can be attained by aggregating all node representations using a readout function like summation or set2set pooling (Vinyals et al., 2015).

2.3 GRAPH DATA AUGMENTATION

It has been shown that the learning performance of GNNs can be improved via graph diffusion, which serves as a homophily-based denoising filter on both features and edges in real graphs (Klicpera et al., 2019). The transformed graphs can also serve as effective augmented views in contrastive learning (Hassani & Khasahmadi, 2020). Inspired by that, we transform a graph G with transition matrix $\boldsymbol{T}$ via graph diffusion and sparsification $\mathbf{S} = \sum_{k=0}^{\infty} \theta_k \boldsymbol{T}^k$ into a new graph with adjacency matrix $\mathbf{S}$ as an augmented view in our framework. While there are many design choices in coefficients θ_k like heat kernel, we employ Personalized PageRank (PPR) with $\theta_k^{PPR} = \alpha(1-\alpha)^k$ due to its superior empirical performance (Hassani & Khasahmadi, 2020). Besides, we randomly remove edges of graphs to attain corrupted graphs as augmented views to validate the robustness of models to different augmentation choices.

3 ITERATIVE GRAPH SELF-DISTILLATION

Intuitively, the goal of contrastive learning on graphs is to learn graph representations that are close in the metric space for positive pairs (graphs with the same labels) and far between negative pairs (graphs with different labels). To achieve this goal, IGSD employ the teacher-student distillation to iteratively refine representations by contrasting latent representations embedded by two networks and using additional predictor and EMA update to avoid collapsing to trivial solutions. Overall, IGSD encourages the closeness of augmented views from the same graph instances while pushing apart the representations from different ones.

3.1 ITERATIVE GRAPH SELF-DISTILLATION FRAMEWORK

In IGSD, we introduce a teacher-student architecture comprises two networks in similar structure composed by encoder f_θ, projector g_θ and predictor h_θ. We denote the components of the teacher network and the student network as $f_{\theta'}$, $g_{\theta'}$ and f_θ, g_θ, h_θ respectively.

The overview of IGSD is illustrated in Figure 1. In IGSD we first augment the original input graphs v_i to get augmented view(s) v_i' and then feed them respectively into two encoders $f_\theta, f_{\theta'}$ for extracting graph representations. The following projectors $g_\theta, g_{\theta'}$ transform graph representations to a higher dimensional projections z, z'. To prevent collapsing into a trivial solution (Grill et al., 2020), a specialized predictor is used in the student network for attaining the prediction $h_\theta(z)$ of the projection z.

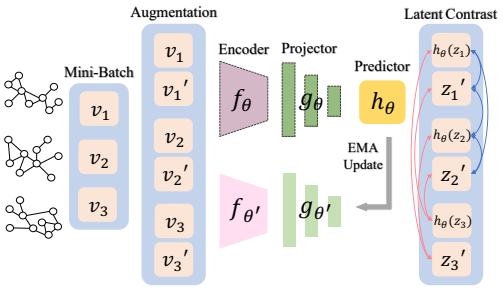


Figure 1: **Overview of IGSD**. Illustration of our framework in the case where we augment input graphs v once to get v' for only one forward pass. Blue and red arrows denote contrast on positive and negative pairs respectively.

To contrast latents $h_\theta(z)$ and z', we use L_2 norm in the latent space to approximate the semantic distance in the input space and the consistency loss can be defined as the mean square error between the normalized prediction $h_\theta(z)$ and projection z'. By passing two instances symmetrically, we can obtain the overall consistency loss:

$$\mathcal{L}^{\text{con}}(G_i, G_i) = \|h_\theta(z_i) - z_i'\|_2^2 + \|h_\theta(z_i') - z_i\|_2^2 \tag{2}$$

With the consistency loss, the teacher network provides a regression target to train the student network, and its parameters θ' are updated as an exponential moving average (EMA) of the student parameters θ after weights of the student model have been updated using gradient descent:

$$\theta_t' \leftarrow \tau\theta_{t-1}' + (1-\tau)\theta_t \tag{3}$$

With the above iterative self-distillation procedure, we can aggregate information for averaging model weights over each training step instead of using the final weights directly.

3.2 UNSUPERVISED LEARNING WITH IGSD

In IGSD, to contrast the anchor G_i with other graph instances G_j (i.e. negative samples), we employ the following unsupervised InfoNCE objective (Oord et al., 2018):

$$\mathcal{L}^{\text{unsup}} = -\mathbb{E}_{G_i \sim \mathcal{G}} \left[\log \frac{\exp\left(\mathcal{L}^{\text{con}}(G_i, G_i)\right)}{\exp\left(\mathcal{L}^{\text{con}}(G_i, G_i)\right) + \sum_{j=1}^{N-1} \mathbb{I}_{i \neq j} \cdot \exp\left(\mathcal{L}^{\text{con}}(G_i, G_j)\right)} \right] \tag{4}$$

At the inference time, as semantic interpolations on samples, labels and latents are effective in obtaining better representations and can improve learning performance greatly (Zhang et al., 2017; Verma et al., 2019; Berthelot et al., 2019), we obtain the graph representation $\tilde{z}$ by interpolating the latent representations $z = f_\theta(G)$ and $z' = f_{\theta'}(G)$ with Mixup function $\mathrm{Mix}_\lambda(\mathrm{a}, \mathrm{b}) = \lambda \cdot \mathrm{a} + (1-\lambda) \cdot \mathrm{b}$:

$$\tilde{z} = \mathrm{Mix}_\lambda(\mathrm{z}, \mathrm{z}') \tag{5}$$

3.3 SEMI-SUPERVISED LEARNING WITH IGSD

To bridge the gap between unsupervised pretraining and downstream tasks, we extend our model to the semi-supervised setting. In this scenario, it is straightforward to plug in the unsupervised loss as a regularizer for representation learning. However, the instance-wise supervision limited to standard supervised learning may lead to biased negative sampling problems (Chuang et al., 2020). To tackle this challenge, we can use a small amount of labeled data further to generalize the similarity loss to handle arbitrary numbers of positive samples belonging to the same class:

$$\mathcal{L}^{\text{sup}} = \sum_{i=1}^{Kl} \frac{1}{K N_{\tilde{\boldsymbol{y}}_i}} \sum_{j=1}^{Kl} \mathbb{I}_{i \neq j} \cdot \mathbb{I}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \mathcal{L}^{\text{con}}(G_i, G_j) \tag{6}$$

where $N_{\tilde{\boldsymbol{y}}_i}$ denotes the total number of samples in the training set that have the same label $\tilde{\boldsymbol{y}}_i$ as anchor i. Thanks to the graph-level contrastive nature of IGSD, we are able to alleviate the biased negative sampling problems (Khosla et al., 2020) with supervised contrastive loss, which is crucial (Chuang et al., 2020) but unachievable in most MI-based contrastive learning models since subgraphs are generally hard to assign labels to. Besides, with this loss we are able to fine-tune our model effectively using self-training where pseudo-labels are assigned iteratively to unlabeled data.

With the standard supervised loss like cross entropy or mean square error $\mathcal{L}(\mathcal{G}_L, \theta)$, the overall objective can be summarized as:

$$\mathcal{L}^{\text{semi}} = \mathcal{L}(\mathcal{G}_L, \theta) + w\mathcal{L}^{\text{unsup}}(\mathcal{G}_L \cup \mathcal{G}_U, \theta) + w'\mathcal{L}^{\text{sup}}(\mathcal{G}_L, \theta) \tag{7}$$

Common semi-supervised learning methods use consistency regularization to measure discrepancy between predictions made on perturbed unlabeled data points to get better prediction stability and generalization performance (Oliver et al., 2018). By contrast, our methods enforce consistency constraints between latents from different views, which acts as a regularizer for learning directly from labels.

Labeled data provides additional supervision about graph classes and alleviates biased negative sampling. However, they are costly to attain in many areas. Therefore, we develop a contrastive self-training algorithm to leverage label information more effectively than cross entropy in the semi-supervised scenario. In the algorithm, we train the model using a small amount of labeled data and then fine-tune it by iterating between assigning pseudo-labels to unlabeled examples and training models using the augmented dataset. In this way, we can harvest massive pseudo-labels for unlabeled examples.

With increasing size of augmented labeled dataset, the discriminative power of IGSD can be improved iteratively by contrasting more positive pairs belonging to the same class. In this way, we accumulate high-quality psuedo-labels after each iteration to compute the supervised contrastive loss in Eq. 6 and make distinction from conventional self-training algorithms (Rosenberg et al., 2005). On the other hand, traditional self-training can only use psuedo-labels for computing cross entropy only.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation Tasks. We conduct experiments by comparing with state-of-the-art models on three tasks. In **graph classification** tasks, we experiment in both the **unsupervised setting** where we only have access to all unlabeled samples in the dataset and the **semi-supervised setting** where we use a small fraction of labeled examples and treat the rest as unlabeled ones by ignoring their labels. In **molecular property prediction** tasks where labels are expensive to obtain, we only consider the **semi-supervised setting**.

Datasets. For graph classification tasks, we employ several widely used benchmark graph kernel datasets (Kersting et al., 2016) for learning and evaluation: 3 bioinformatics datasets (MUTAG, PTC_MR, NCI1) and 3 social network datasets (COLLAB, IMDB-BINARY, IMDB-MULTI) with statistics summarized in Table 1.

In the semi-supervised graph regression tasks, we use the QM9 dataset containing 134,000 drug-like organic molecules (Ramakrishnan et al., 2014) and select the first ten physicochemical properties as regression targets for training and evaluation.

Baselines. In the unsupervised graph classification, we compare with the following representative baselines: CMC-Graph (Hassani & Khasahmadi, 2020), InfoGraph (Sun et al., 2019), Graph2Vec (Narayanan et al., 2017) and Graph Kernels including Random Walk Kernel (Gärtner et al., 2003), Shortest Path Kernel (Kashima et al., 2003), Graphlet Kernel (Shervashidze et al., 2009), Weisfeiler-Lehman Sub-tree Kernel (WL SubTree) (Shervashidze et al., 2011), Deep Graph Kernels (Yanardag & Vishwanathan, 2015) and Multi-Scale Laplacian Kernel (MLG) (Kondor & Pan, 2016).

For the semi-supervised graph classification, we compare our method with competitive baselines like InfoGraph, InfoGraph* and Mean Teachers. And the GIN baseline doesn't have access to the unlabeled data. In the semi-supervised molecular property prediction tasks, baselines include InfoGraph, InfoGraph* and Mean Teachers (Tarvainen & Valpola, 2017).

Model Configuration. In our framework, We use GCNs (Kipf & Welling, 2016) and GINs (Xu et al., 2018) as encoders to attain node representations for the unsupervised and semi-supervised graph classification respectively. For semi-supervised molecular property prediction tasks, we employ message passing neural networks (MPNNs) (Gilmer et al., 2017) as our backbone encoders to encode molecular graphs with rich edge attributes. All projectors and predictors are implemented as two-layer MLPs.

In semi-supervised molecular property prediction tasks, we generate multiple views based on edge attributes (bond types) of rich-annotated molecular graphs for improving performance. Specifically, we perform label-preserving augmentation to attain multiple diffusion matrixes of every graph on different edge attributes while ignoring others respectively. The diffusion matrix gives a denser graph based on each type of edges to leverage edge features better. We train our models using different numbers of augmented training data and select the amount using cross validation.

For parameter tuning, we select number of GCN layers over $\{2, 8, 12\}$, batch size over $\{16, 32, 64, 128, 256, 512\}$, number of epochs over $\{20, 40, 100\}$ and learning rate over $\{1e\text{-}4, 1e\text{-}3\}$ in unsupervised graph classification. The hyper-parameters we tune for semi-supervised graph classification and molecular property prediction are the same in (Xu et al., 2018) and (Sun et al., 2019), respectively. In all experiments, we set the weighting coefficient of Mixup function to be 0.5 and tune our projection hidden size over $\{1024, 2048\}$ and projection size over $\{256, 512\}$ since bigger projection heads are able to improve representation learning (Chen et al., 2020b). We start self-training after 30 epochs and tune the number of iterations over $\{20, 50\}$, pseudo-labeling threshold over $\{0.9, 0.95\}$.

For unsupervised graph classification, we adopt LIB-SVM (Chang & Lin, 2011) with C parameter selected in $\{1e\text{-}3, 1e\text{-}2, \ldots, 1e2, 1e3\}$ as our downstream classifier. Then we use 10-fold cross validation accuracy as the classification performance and repeat the experiments 5 times to report the mean and standard deviation. For semi-supervised graph classification, we randomly select 5% of training data as labeled data and the rest is treated as unlabeled one and report the best test set accuracy in 300 epochs. Following the experimental setup in (Sun et al., 2019), we randomly choose

	Datasets	MUTAG	IMDB-B	IMDB-M	NCI1	COLLAB	PTC
Datasets	# graphs	188	1000	1500	4110	5000	344
	# classes	2	2	3	2	3	2
	Avg # nodes	17.9	19.8	13.0	29.8	74.5	25.5
Graph Kernels	Random Walk	83.7 ± 1.5	50.7 ± 0.3	34.7 ± 0.2	OMR	OMR	57.9 ± 1.3
	SHORTEST PATH	85.2 ± 2.4	55.6 ± 0.2	38.0 ± 0.3	51.3 ± 0.6	49.8 ± 1.2	58.2 ± 2.4
	GRAPHLET KERNEL	81.7 ± 2.1	65.9 ± 1.0	43.9 ± 0.4	53.9 ± 0.4	56.3 ± 0.6	57.3 ± 1.4
	WL SUBTREE	80.7 ± 3.0	72.3 ± 3.4	47.0 ± 0.5	55.1 ± 1.6	50.2 ± 0.9	58.0 ± 0.5
	DEEP GRAPH	87.4 ± 2.7	67.0 ± 0.6	44.6 ± 0.5	54.5 ± 1.2	52.1 ± 1.0	60.1 ± 2.6
	MLG	87.9 ± 1.6	66.6 ± 0.3	41.2 ± 0.0	>1 Day	>1 Day	$\mathbf{63.3 \pm 1.5}$
Unsupervised	GRAPH2VEC	83.2 ± 9.6	71.1 ± 0.5	50.4 ± 0.9	73.2 ± 1.8	47.9 ± 0.3	60.2 ± 6.9
	INFOGRAPH	89.0 ± 1.1	74.2 ± 0.7	49.7 ± 0.5	73.8 ± 0.7	67.6 ± 1.2	61.7 ± 1.7
	CMC-GRAPH	89.7 ± 1.1	74.2 ± 0.7	51.2 ± 0.5	75.0 ± 0.7	68.9 ± 1.9	62.5 ± 1.7
	OURS (RANDOM)	85.7 ± 2.1	71.6 ± 1.2	49.2 ± 0.6	75.1 ± 0.4	65.8 ± 1.0	57.6 ± 1.5
	OURS	$\mathbf{90.2 \pm 0.7}$	$\mathbf{74.7 \pm 0.6}$	$\mathbf{51.5 \pm 0.3}$	$\mathbf{75.4 \pm 0.3}$	$\mathbf{70.4 \pm 1.1}$	61.4 ± 1.7

Table 1: **Graph classification accuracies (%) for kernels and unsupervised methods on 6 datasets.** We report the mean and standard deviation of final results with five runs. '>1 day' represents that the computation exceeds 24 hours. 'OMR' means out of memory error.

Datasets	IMDB-B	IMDB-M	COLLAB	NCI1
MEAN TEACHERS	69.0	49.3	72.5	71.1
INFOGRAPH*	71.0	49.3	67.6	71.1
GIN (SUPERVISED ONLY)	67.0	50.0	71.4	67.9
OURS (UNSUP)	72.0	50.0	72.6	70.6
OURS (SUPCON)	75.0	52.0	73.4	67.9
GIN (SUPERVISED ONLY)+SELF-TRAINING	72.0	51.3	70.4	74.0
OURS (UNSUP)+SELF-TRAINING	73.0	54,0	71	72.5
OURS (SUPCON)+SELF-TRAINING	**77.0**	**55.3**	**73.6**	**77.1**

Table 2: **Graph classification accuracies (%) of semi-supervised experiments on 4 datasets.** We report the best results on test set in 300 epochs.

5000, 10000, 10000 samples for training, validation and testing respectively and the rest are treated as unlabeled training data for the molecular property prediction tasks.

4.2 NUMERICAL RESULTS

Results on unsupervised graph classification. We first present the results of the unsupervised setting in Table 1. All graph kernels give inferior performance except in the PTC dataset. The Random Walk kernel runs out of memory and the Multi-Scale Laplacian Kernel suffers from a long running time (exceeds 24 hours) in two larger datasets. IGSD outperforms state-of-the-art baselines like InfoGraph and CMC-Graph, showing that IGSD can learn expressive graph-level representations for downstream classifiers. Besides, our model still achieve competitive results in datasets like IMDB-M and NCI1 with random dropping augmentation, which demonstrates the robustness of IGSD with different choices of data augmentation strategies.

Results on semi-supervised graph classification. We further apply our model to semi-supervised graph classification tasks with results demonstrated in Table 2, where we set w and w' in Eq. 7 to be 1 and 0 as Ours (Unsup) while 0 and 1 as Ours (SupCon). In this setting, our model performs better than Mean Teachers and InfoGraph*. Both the unsupervised loss and supervised contrastive loss provide extra performance gain compared with GIN using supervised data only. Besides, both of their performance can be improved significantly combined using self-training especially with supervised contrastive loss. It makes empirical sense since self-training iteratively assigns psuedo-labels with high confidence to unlabeled data, which provides extra supervision on their categories under contrastive learning framework.

Results on semi-supervised molecular property prediction. We present the regression performance of our model measured in the QM9 dataset in Figure 2. We display the performance of our model and baselines as mean square error ratio with respect to supervised results and our model outperforms all baselines in 9 out of 10 tasks compared with strong baselines InfoGraph, InfoGraph* and Mean Teachers. And in some tasks like R2 (5), U0 (7) and U (8), DGSI achieves significant

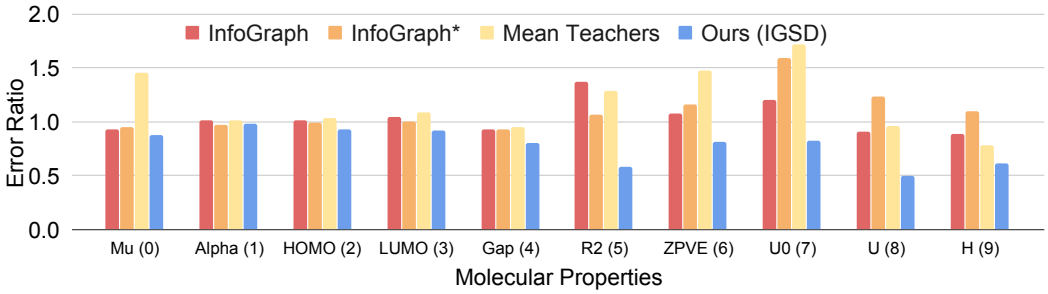
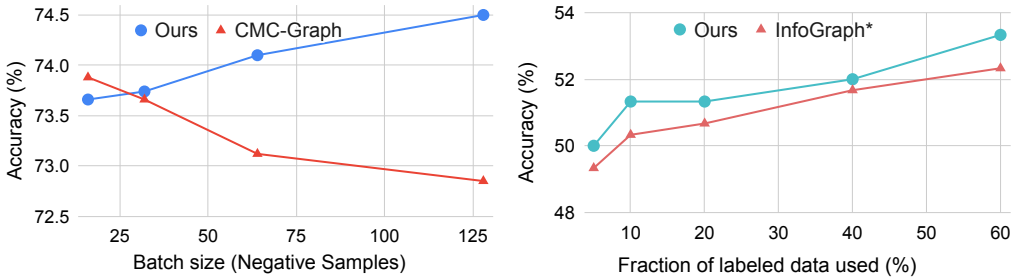


Figure 2: **Semi-supervised molecular property prediction results in terms of mean absolute error (MAE) ratio.** The histogram shows error ratio with respect to supervised results (1.0) of every semi-supervised models. Lower scores are better and a model outperforms the supervised baseline when the score is less than 1.0.



(a) (b)

Figure 3: **Ablation studies** (a) Unsupervised learning performance with different batch size (different amount of negative pairs); (b) Semi-supervised graph classification accuracy with different proportion of labeled data.

performance gains against its counterparts, which demonstrates the ability to transfer knowledge learned from unsupervised data for supervised tasks.

4.3 Ablation Studies and Analysis

Performance with self-training. We first investigate the effects of self-training for our model performance in table 2. Results show that self-training can improve the GIN baseline and our models with unsupervised loss (Unsup) or supervised contrastive loss (SupCon). The improvement is even more significant combined with supervised contrastive loss since high-quality pseudo-labels provide additional information of graph categories. Moreover, our self-training algorithm consistently outperforms the traditional self-training baseline, which further validates the superiority of our model.

Performance with different amount of negative pairs. We then conduct ablation experiments on the amount of negative pairs by varying batch size over $\{16, 32, 64, 128\}$ with results on IMDB-BINARY dataset shown in Figure 3a. Both methods contrast negative pairs batch-wise and increasing batch size improves the performance of IGSD while degrades CMC-Graph. When batch size is greater than 32, IGSD outperforms CMC-Graph and the performance gap becomes larger as the batch size increases.

Performance with different proportion of labeled data. We also investigate the performance of different models with different proportion of labeled data with IMDB-BINARY dataset. As illustrated in Figure 3b, IGSD outperforms strong InfoGraph* baseline given different amount of labeled data consistently. And the performance gain is most significant when the fraction of labeled data is 10% since our models can leverage labels more effectively by regularizing original unsupervised learning objective when labels are scarce.

5 Related Work

Contrastive Learning Modern unsupervised learning in the form of contrastive learning can be categorized into two types: context-instance contrast and context-context contrast (Liu et al., 2020). The context-instance contrast, or so-called global-local contrast focuses on modeling the belonging relationship between the local feature of a sample and its global context representation. Most unsupervised learning models on graphs like DGI (Veličković et al., 2018), InfoGraph (Sun et al., 2019), CMC-Graph (Hassani & Khasahmadi, 2020) fall into this category, following the *InfoMax principle* to maximize the the mutual information (MI) between the input and its representation. However, estimating MI is notoriously hard in MI-based contrastive learning and in practice tractable lower bound on this quantity is maximized instead. And maximizing tighter bounds on MI can result in worse representations without stronger inductive biases in sampling strategies, encoder architecture and parametrization of MI estimators (Tschannen et al., 2019). Besides, the intricacies of negative sampling in MI-based approaches impose key research challenges like improper amount of negative samples or biased negative sampling (Tschannen et al., 2019; Chuang et al., 2020). Another line of contrastive learning approaches called context-context contrast directly study the relationships between the global representations of different samples as what metric learning does. For instance, a recently proposed model BYOL (Grill et al., 2020) bootstraps the representations of the whole images directly. Focusing on global representations between samples and corresponding augmented views also allows instance-level supervision to be incorporated naturally like introducing supervised contrastive loss (Khosla et al., 2020) into the framework for learning powerful representations. Graph Contrastive Coding (GCC) (Qiu et al., 2020) is a pioneer to leverage instance discrimination as the pretext task for structural information pre-training. However, our work is fundamentally different from theirs. GCC focuses on structural similarity with InfoNCE as learning objective to find common and transferable structural patterns across different graph datasets and the contrastive scheme is done through subgraph instance discrimination. On the contrary, our model aims at learning graph-level representation by directly contrasting graph instances such that data augmentation strategies and graph labels can be utilized naturally and effectively.

Knowledge Distillation Knowledge distillation (Hinton et al., 2015) is a method for transferring knowledge from one architecture to another, allowing model compression and inductive biases transfer. Self-distillation (Furlanello et al., 2018) is a special case when two architectures are identical, which can iteratively modify regularization and reduce over-fitting if perform suitable rounds (Mobahi et al., 2020). However, they often focus on closing the gap between the predictive results of student and teacher rather than defining similarity loss in latent space for contrastive learning.

Semi-supervised Learning Modern semi-supervised learning can be categorized into two kinds: multi-task learning and consistency training between two separate networks. Most widely used semi-supervised learning methods take the form of multi-task learning: $\arg\min_\theta \mathcal{L}_l(D_l, \theta) + w\mathcal{L}_u(D_u, \theta)$ on labeled data D_l and unlabeled data D_u. By regularizing the learning process with unlabeled data, the decision boundary becomes more plausible. Another mainstream of semi-supervised learning lies in introducing student network and teacher network and enforcing consistency between them (Tarvainen & Valpola, 2017; Miyato et al., 2019; Lee, 2013). It has been shown that semi-supervised learning performance can be greatly improved via unsupervised pre-training of a (big) model, supervised fine-tuning on a few labeled examples, and distillation with unlabeled examples for refining and transferring the task-specific knowledge (Chen et al., 2020b). However, whether task-agnostic self-distillation would benefit semi-supervised learning is still underexplored.

6 Conclusions

In this paper, we propose IGSD, a novel unsupervised graph-level representation learning framework via self-distillation. Our framework iteratively performs teach-student distillation by contrasting augmented views of graph instances. Experimental results in both unsupervised and semi-supervised settings show that IGSD is not only able to learn effective graph representations competitive with state-of-the-art models but also robust with choices of encoders and augmentation strategies. In the future, we plan to apply our framework to other graph learning tasks and investigate the design of view generators to generative effective views automatically.

REFERENCES

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pp. 585–591, 2002.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning, 2020.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pp. 2224–2232, 2015.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks, 2018.

Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning theory and kernel machines*, pp. 129–143. Springer, 2003.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pp. 1024–1034, 2017.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of International Conference on Machine Learning*, pp. 3451–3461. 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. doi: 10.1109/cvpr42600.2020.00975. URL http://dx.doi.org/10.1109/cvpr42600.2020.00975.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks, 2019.

Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 321–328, 2003.

Kristian Kersting, Nils M. Kriege, Christopher Morris, Petra Mutzel, and Marion Neumann. Benchmark data sets for graph kernels, 2016. URL http://graphkernels.cs.tu-dortmund.de.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Johannes Klicpera, Stefan Weißenberger, and Stephan Günnemann. Diffusion improves graph learning, 2019.

Risi Kondor and Horace Pan. The multiscale laplacian graph kernel. In *Advances in Neural Information Processing Systems*, pp. 2990–2998, 2016.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive, 2020.

Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, Aug 2019. ISSN 1939-3539. doi: 10.1109/tpami.2018.2858821. URL http://dx.doi.org/10.1109/TPAMI.2018.2858821.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L Bartlett. Self-distillation amplifies regularization in hilbert space. *arXiv preprint arXiv:2002.05715*, 2020.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005*, 2017.

Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin D. Cubuk, and Ian J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Aug 2020. doi: 10.1145/3394486.3403168. URL http://dx.doi.org/10.1145/3394486.3403168.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pp. 488–495, 2009.

Nino Shervashidze, Pascal Schweitzer, Erik Jan Van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(9), 2011.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, 2017.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447. PMLR, 2019.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets, 2015.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374, 2015.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When does self-supervision help graph convolutional networks? *arXiv preprint arXiv:2006.09136*, 2020.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization, 2017.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. Data augmentation for graph neural networks, 2020.

Yadi Zhou, Fei Wang, Jian Tang, Ruth Nussinov, and Feixiong Cheng. Artificial intelligence in covid-19 drug repurposing. *The Lancet Digital Health*, 2020.