Multi-Scale Distillation from Multiple Graph Neural Networks

Chunhai Zhang[1], **Jie Liu**[1,2*], **Kai Dang**[1], **Wenzheng Zhang**[1]

[1]College Of Artificial Intelligence, Nankai University, Tianjin, China
[2]Cloopen AI Research, Beijing, China
{chzhang, dangkai, wzzhang}@mail.nankai.edu.cn, jliu@nankai.edu.cn

Abstract

Knowledge Distillation (KD), which is an effective model compression and acceleration technique, has been successfully applied to graph neural networks (GNNs) recently. Existing approaches utilize a single GNN model as the teacher to distill knowledge. However, we notice that GNN models with different number of layers demonstrate different classification abilities on nodes with different degrees. On the one hand, for nodes with high degrees, their local structures are dense and complex, hence more message passing is needed. Therefore, GNN models with more layers perform better. On the other hand, for nodes with low degrees, whose local structures are relatively sparse and simple, the repeated message passing can easily lead to over-smoothing. Thus, GNN models with less layers are more suitable. However, existing single-teacher GNN knowledge distillation approaches which are based on a single GNN model, are sub-optimal. To this end, we propose a novel approach to distill multi-scale knowledge, which learns from multiple GNN teacher models with different number of layers to capture the topological semantic at different scales. Instead of learning from the teacher models equally, the proposed method automatically assigns proper weights for each teacher model via an attention mechanism which enables the student to select teachers for different local structures. Extensive experiments are conducted to evaluate the proposed method on four public datasets. The experimental results demonstrate the superiority of our proposed method over state-of-the-art methods. Our code is publicly available at https://github.com/NKU-IIPLab/MSKD.

Introduction

Many large-scale deep models have been put forward and achieved significant success in many fields. Nonetheless, the huge computational complexity and massive storage requirements restrict their deployments when the computation resource is limited. To address this problem, knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015; Romero et al. 2014; Lan, Zhu, and Gong 2018; Zhou et al. 2021) has been widely investigated. It is one of the main streams of model compression and acceleration technique (Wu et al. 2016; Courbariaux, Bengio, and David 2015; Yu et al. 2017; Zhai et al. 2016; Hinton, Vinyals, and Dean 2015). By transferring the knowledge of a cumbersome teacher model to a compact student model, the student is able to master the expertise of the teacher and can be readily deployed.

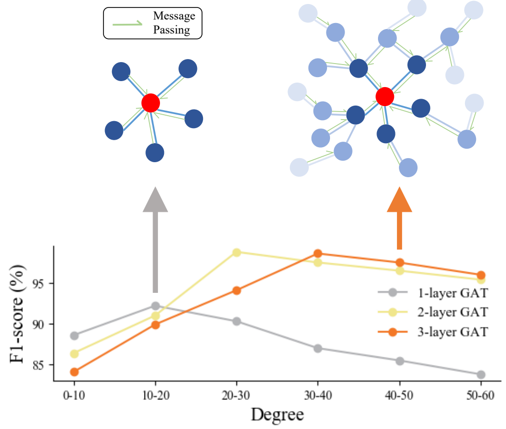


Figure 1: The performance of l-Layer GNN models on different Local Structures. Using PPI dataset as an illustrative example, the line chart shows each model's node classification F1-score on subsets of node with certain range of degree.

Recently, KD has been successfully generalized to graph neural networks (GNNs) (Yang et al. 2020; Yao et al. 2020; Jing et al. 2021). Most existing approaches utilize a single GNN model as the teacher to distill knowledge. However, we notice that the node classification ability of GNN models with l hidden layers (l-layer GNN) varies on the nodes with different degrees, as shown in figure 1. On the one hand, l-layer GNN models with more layers perform better on nodes with high degrees, because their local structures are dense and complex, where hence more message passing is helpful. On the other hand, l-layer GNN models with less layers are more suitable for nodes with low degrees, since their local structures are relatively sparse and simple. The repeated message passing can easily lead to over-smoothing for such local structures. Therefore, the existing single-teacher knowledge distillation approaches are sub-optimal when dealing with graphs consisting of various local structures.

[*]Corresponding author.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this paper, we propose a **M**ulti-**S**cale **K**nowledge **D**istillation (MSKD) approach to exploit multi-scale local structure knowledge by amalgamating multiple l-layer GNN teacher models. The proposed approach enables the student model to encode multi-scale topological structure semantics from the teacher models, and hence achieve ideal classification ability on various local structures with different sparsity which is the degree of the core node.

Based on the previous analysis in Figure 1, the multiple teachers set in MSKD consist of GNN models with different number of layers which can cover a wide range of node degrees on the graph. To cover enough local structure information, we analyze the characteristics of GNN with different number of layers and propose a principle of teacher model set construction based on the GNN model degradation (Zhang et al. 2021). Then we employ Topological Semantic Mapping to effectively transfer structure information from each teacher model. More important, we devise an attention mechanism in MSKD to adaptively assign weights for the GNN teacher models. Learning from GNN models with different number of layers via an attention mechanism rather than from a single GNN model can avoid considering node degrees only in a fixed scope. By taking advantage of multi-scale degree-related knowledge transfer from multiple teachers, the student can be effectively optimized with more comprehensive guidance. Moreover, the proposed MSKD loss can be combined with the classic knowledge amalgamation (KA) (Shen et al. 2019; Luo et al. 2019; Jing et al. 2021), which aims to learn a student network from multiple teachers of different domains, to further improve the distillation performance.

To validate the effectiveness of the proposed method, we conduct extensive experiments on four different datasets, including protein-protein interaction (PPI) dataset, Cora, CiteSeer and PubMed. PPI is used for the multi-label node classification task, and the last three citation networks are used for the single-label node classification task. Experimental results demonstrate the superiority of MSKD over state-of-the-art methods.

Our contributions overall are summarized as following.

- We propose a novel approach to effectively learn multi-scale topological semantics from multiple GNN teacher models. To the best of our knowledge, this is the first work of multiple GNN Knowledge Distillation from the perspective of multi-scale structure information.
- We devise an attentional amalgamation technique which is selective about distilling effective local structure knowledge from multiple GNN teacher models.
- Extensive experiments are conducted to evaluate the proposed method on four different datasets. The results illustrate that our model outperforms the state-of-the-art approaches.

Related Work

Graph Neural Networks

Graph neural networks (Kipf and Welling 2017; Velickovic et al. 2017; Xu et al. 2018; Lei et al. 2020), which can effectively combine the expressive power of graphs with deep learning, have achieved unprecedented advances in recent years. Among them, graph convolutional networks (GCNs) (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017) can be seen as some enhanced GNN models. Specifically, (Kipf and Welling 2017) proposed a scalable approach on graph-structure data based on a localized first-order approximation of spectral graph convolutions. This is the first job that generalizes convolutional neural networks on the graph. Graph attention network (GAT) (Velickovic et al. 2017), which leverages masked self-attentional layers to address the weaknesses of the earlier GCNs, and assigns different weights to different nodes in a neighborhood by stacking layers. Deep graph infomax (DGI) (Veličković et al. 2018) maximizing mutual information between patch representations and corresponding high-level summaries of graphs. Furthermore, graph isomorphism network (GIN) (Xu et al. 2018) is a simple architecture that is provably the most expressive among the class of GNNs and as powerful as the Weisfeiler-Lehman graph isomorphism test. We utilize GAT (Velickovic et al. 2017) and GCN (Kipf and Welling 2017) in our model to validate the effectiveness and generalization of our model on GNNs.

Knowledge Distillation

Knowledge distillation (Hinton, Vinyals, and Dean 2015; Romero et al. 2014; Zagoruyko and Komodakis 2016) is one of the main streams of model compression and acceleration techniques (Wu et al. 2016; Courbariaux, Bengio, and David 2015; Yu et al. 2017; Zhai et al. 2016; Hinton, Vinyals, and Dean 2015). It effectively learns a small, thin student from a large, cumbersome teacher so that the student can hold a similar performance as the teacher's. (Hinton, Vinyals, and Dean 2015) first formally present the concept of knowledge distillation and use the logits of a large deep model as the teacher knowledge to help improve the performance of the student. FitNet (Romero et al. 2014) uses not only the outputs but also the intermediate representations as the knowledge. It adds an additional fully connected layer to match the features of teacher and student. (Zagoruyko and Komodakis 2016) proposes several methods of transferring attention to force the student to mimic the attention maps of the teacher. SemCKD (Chen et al. 2021) proposes a method that automatically assigns proper target layers of the teacher for each student layer with an attention mechanism. Moreover, (Yang et al. 2020) proposes a local structure preserving module to extract non-grid data, i.e., the local structure information as distributions and minimizes the distance between the distributions of the teacher and the student to improve the performance of distillation. They all distill knowledge from a single teacher which is one-sided for training a student.

Knowledge Amalgamation

Different teachers contain different useful information for students. Thus, many multi-teacher knowledge distillation models have been proposed (Sau and Balasubramanian 2016; You et al. 2017) to utilize the knowledge from different teachers to train the student. Among them, knowledge amalgamation (Shen et al. 2019; Luo et al. 2019; Jing et al. 2021) trains a student by transferring the knowledge

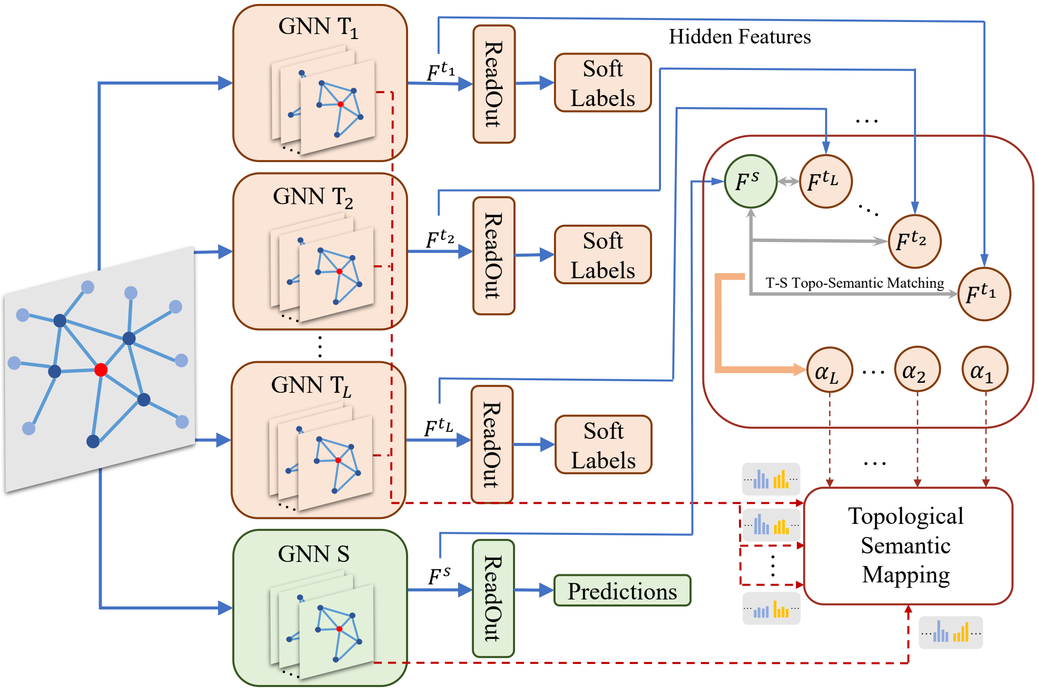


Figure 2: An overview of the proposed Multi-scale Knowledge Distillation (MSKD). GNN T$_1$, GNN T$_2$, GNN T$_L$ and GNN S represent the graph neural networks from the 1^{st}, 2^{nd}, L^{th} pre-trained teachers and the compact student. F^{t_1}, F^{t_2}, F^{t_L} and F^s are the hidden features of the 1^{st}, 2^{nd}, L^{th} teachers and the student, respectively. T-S means the pairs of the teachers and the student. The attention mechanism, which is shown in the right pane, adaptively assigns topological semantic-related weights for the teachers to help improve the distillation performance.

of multiple teachers from different domains. (Shen et al. 2019) is the first KA work that learns a student model by using multiple teachers which focus on different classification problems. Thus, the student can accomplish a comprehensive classification task. (Jing et al. 2021) proposes a slimmable graph convolutional operation to learn varying-dimension features from teachers, and a topological attribution map (TAM) scheme to learn a student from multiple teachers' topological semantics. This is currently the only multi-teacher KD job that applies to GNNs. However, the teachers contribute equally in these methods which ignore the importance of the different knowledge. Our model is devised to be a multi-teacher KD model which is able to fully learn the multi-scale knowledge on the graph. And an attention mechanism is devised in our method to address the above problem. Our method can also be seen as KA because we amalgamate the knowledge from different scales of node degrees. And the proposed method can be easily combined with classic KA by integrating the local knowledge learned by each teacher's different-layer-versions to further improve the distillation performance.

Problem Formulation

In this work, we aim to learn a compact student GNN model that amalgamates multi-scale topological knowledge from multiple l-layer GNN teacher models. Given an unweighted and undirected graph $G = (\mathcal{V}, \mathcal{E}, X)$, where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ is the node set, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, and $X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^{n \times d}$ is the feature matrix, x_i stands for the feature vector of node v_i, L teacher models $\mathcal{T} = \{t_1, t_2, ..., t_L\}$ are pre-trained. Each teacher model t_l is a GNN model with l hidden layers. As previously analyzed, each teacher model has its own preference on the scale of topological structure. So, the goal of the proposed task is to learn a compact student model s that adaptively transfer proper knowledge from each teacher model.

Method

In this section, we present the method of Multi-Scale Knowledge Distillation (MSKD) from multiple GNN models, whose overall framework is shown in Figure 2. We first introduce the construction of the set of pre-trained l-layer GNN teacher models. In what follows, the topological knowledge distillation from each l-layer GNN teacher model is described. Then we devise an attention multi-scale topological semantics amalgamation module to adaptively distill proper knowledge from the teacher models. Finally, the overall loss function and algorithm are given.

Multiple l-layer GNN Teacher Models

The teacher models in MSKD are l-layer GNN models which learn knowledge at different scales of topological semantic. For an l-layer GNN, a neighborhood aggregation scheme is performed to capture the l-hop information surrounding each node, which makes the GNN model encodes the topological semantic information at the scale of l-hop. As aforementioned, the higher the node degree is, the more hidden layer is needed to encode the complex structure information.

Without losing generality, we choose two kinds of representative GNN models, i.e. GCN and GAT, to validate the effectiveness and generalization of the proposed method. They can be formulated as:

$$h^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}h^{(l)}W^{(l)}), \tag{1}$$

$$h'_i = \sigma(\frac{1}{C}\sum_{c=1}^{C}\sum_{j\in\mathcal{N}_i}\omega_{ij}^{c}W^{c}h_j), \tag{2}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix A with added self-connections, I_N is the identity matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, $h^{(l)}$ is the representation of the l^{th} GCN, $W^{(l)}$ is the weight matrix in the l^{th} layer, $\sigma(\cdot)$ is the activation function. In Equation (2), C is the number of attention mechanisms in GAT, $\mathcal{N}_i$ is the neighborhood of node v_i, ω^c is the c^{th} attention coefficient, W^c is the corresponding input linear transformation's weight matrix, h_j is the representation of the node v_j.

Furthermore, it is important to determine the proper size of the teacher set $\mathcal{T} = \{t_l\}_{l=1}^{L}$, where L corresponds to both the value number of GNN teacher models and the number of hidden layers of the GNN teacher model t_L. Hence, the value of l ranges from 1 to L. Here, we should choose a proper L to balance the topological structure scale coverage and the number of the pre-trained teacher models. (Zhang et al. 2021) shows that deep GNN models tend to exhibit a model degradation problem with the increase of the number of hidden layers number due to the over-smoothed node representations. Based on this phenomenon, we pre-train l-layer GNN teacher models with increasing l at the step of 1, until there is an obvious performance decrease tendency.

Topological Semantic Mapping

To effectively distill knowledge from each GNN teacher model, we seek to transfer the topological semantics captured by the teacher models. We propose to preserve topological information by utilizing local structures, which is corresponding to a subgraph formed by a center node x_i and its neighbors $\mathcal{N}_i$. As suggested by (Yang et al. 2020), the local structures learned by a teacher model t_l can be denoted as a set of vectors $\mathrm{LS}^{t_l} = \{LS_1^{t_l}, LS_2^{t_l}, ..., LS_n^{t_l}\}$, where $LS_i^{t_l} \in \mathbb{R}^{\mathcal{D}_i}$, $\mathcal{D}_i$ is the degree of the local structure $LS_i^{t_l}$'s center node v_i. Each element of the vector can be calculated as

$$LS_{ij}^{t_l} = \mathrm{softmax}(\exp(-\frac{1}{2}||h_i - h_j||^2)), \tag{3}$$

where the function in softmax measures the similarity of the given pair of nodes v_i and v_j.

As implies in Equation (1) and (2), GNNs learn topological semantics by passing and aggregating messages based on the local structures. Therefore, we employ Topological Semantic Mapping to measure the distance of the local structure semantic pair of the student LS^s and the target teacher LS^{t_l} by Kullback-Leibler (KL) divergence:

$$Dist(\mathrm{LS}^s, \mathrm{LS}^{t_l}) = \mathcal{L}_{KL}(\mathrm{LS}^s, \mathrm{LS}^{t_l}) \tag{4}$$

where t_l is the l^{th} teacher. By minimizing the distances between local structure semantic pairs, the topological semantics of the teachers, which contains multi-scale degree-related knowledge, can be transferred to the student.

Attentional Topological Semantic Amalgamation

As shown in Figure 1, each teacher's performance varies according to the node degree. The higher the degree of v is, the more message passing is needed for modeling the complex local structure. That is why GNN models with more hidden layers perform better on the nodes with a high degree. And vise versa.

Since GNN teacher models with different number of layers contain different topological semantics, it is desirable to adaptively distill knowledge from teachers. To improve the distillation performance, the student should learn weighted multi-scale knowledge from the teachers with different number of layers according to the matching degree of each teacher-student pair. We separately project the hidden features of the student $F^s \in \mathbb{R}^K$ and each teacher $F^{t_m} \in \mathbb{R}^K$ into two subspaces via MLP (Vaswani et al. 2017) which aims to deal with the problem of noise and sparseness, where K is the categories. Then, the attention weight of the l^{th} teacher α_l is calculated as follows:

$$\alpha_l = \frac{MLP(F^s)^T MLP(F^{t_l})}{\sum_{l=1}^{L}(MLP(F^s)^T MLP(F^{t_l}))}, l = 1, 2, ...L, \tag{5}$$

where L is the number of teachers, the corresponding weights satisfy $\sum_{l=1}^{L}\alpha_l = 1$, and we use logits $g \in \mathbb{R}^K$ as the hidden features in this paper. Attention-based allocation reduces the loss of the student in distilling multi-scale degree-related knowledge from multiple teachers.

With the attention allocation and the Topological Semantic Mapping in Equation (4), we can obtain the multi-scale knowledge distilling loss as follows:

$$\begin{aligned}\mathcal{L}_{MSKD} &= \sum_{l=1}^{L}\alpha_l Dist(\mathrm{LS}^s, \mathrm{LS}^{t_l}) \\ &= \frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N}\sum_{j:(j,i)\in\mathcal{E}}\alpha_l LS_{ij}^{s} log(\frac{LS_{ij}^{s}}{LS_{ij}^{t_l}}).\end{aligned} \tag{6}$$

With the help of the learned attention distributions, the MSKD loss is denoted as the weighted summation of each KL divergence between the student and each teacher.

Loss Function

We design joint loss functions by combining classic response-based loss and MSKD loss.

Response-based knowledge distillation transfers the knowledge via logits. The predicted probabilities $p = \sigma(g/T)$, where temperature T is a hyperparameter and the soft labels can be softer with a higher T. We set T to 4 in this paper for a fair comparison. For classification tasks, the

Algorithm 1: Multi-scale Knowledge Distillation.

```
Input: Training graph $G$; Pre-trained teacher GNNs $\mathcal{T} = \{t_l\}_{l=1}^{L}$ and their parameters $\theta^{t_1}, \theta^{t_2}, ..., \theta^{t_l}$; A student model with randomly initialized parameter $\theta^s$;
Output: A compact student model with a broad view of node degrees;
while $\theta^s$ is not converged do
    By feeding $G$ into $\theta^{t_l}$ and $\theta^s$, obtain local structure semantic representations $\mathrm{LS}^{t_l}$ and $\mathrm{LS}^s$.
    Compute the soft labels $p^{t_l}$ and the prediction $p^s$ from the outputs of teacher $t_l$ and the student $s$.
    Utilize attention mechanism as Equation (5).
    By backward propagating the loss (in Equation (8)) gradients to update parameters $\theta^s$.
end while
```

loss function is devised to be a cross entropy loss (CE) combines the minimization of KL divergence between p^s and soft target p^{t_l} of each teacher model:

$$\mathcal{L}_{KD} = \mathcal{L}_{CE}(\sigma(g^s), y) + \sum_{l=1}^{L} T^2 \mathcal{L}_{KL}(p^s, p^{t_l}), \quad (7)$$

where y is the one-hot label.

By further introducing the MSKD loss in Equation (6), we can obtain the overall loss

$$\mathcal{L} = \mathcal{L}_{KD} + \lambda \mathcal{L}_{MSKD}, \quad (8)$$

where the hyperparameter λ is used to balance two individual loss terms. The learning procedure is summarized in Algorithm 1.

Experiment

In this section, we first introduce our datasets and the comparison methods. Then, we provide our experimental settings. We show the comparison of the F1-scores of our method and baselines to demonstrate the superiority of the proposed method. In addition, we also provide results to explain and support the success of our multi-scale knowledge distillation strategy in helping student models learn as much multi-scale local structure knowledge as possible through well-designed experiments. Moreover, ablation studies and visualization experiments are also conducted.

Datasets

We conduct a series of node classification tasks on four different datasets, i.e., PPI (Zitnik and Leskovec 2017), Cora, CiteSeer and PubMed (Sen et al. 2008). PPI is used for the multi-label node classification task, and the last three citation networks are used for the single-label node classification task.

- PPI contains 24 graphs that come from different human tissues and 121 categories, where 20 graphs are used for training, 2 graphs are used for validating and the left 2 graphs are used for testing. The average number of the nodes of each graph is 2372 and the average node degree is 29.32. The dimension of the input feature is 50.
- Cora, CiteSeer and PubMed are three citation network datasets. They contain 7, 6 and 3 categories. The edges are made up of the links between two papers. These datasets contain 2708, 3327, 19,717 nodes, and 10,556, 9104, 88,648 edges. Thus, the average node degrees are 3.90, 2.74 and 4.50. The dimensions of the input feature are 1433, 3703 and 500.

Comparison Methods

To demonstrate the performance of our proposed method, we compare it with various knowledge distillation models. Details about these baselines are as follows:

- FitNet. FitNet (Romero et al. 2014) uses the outputs and the intermediate representations learned by the teachers to train a student. Besides, it introduces an additional mapping function to calculate the L_2 distance between the features of the teacher and the mapped student.
- AT. Attention Transfer (AT) (Zagoruyko and Komodakis 2016) is another KD model that utilizes the intermediate representations. It proposes several attention transfer methods to force the student to mimic the attention maps of a powerful teacher network.
- LSP. LSP (Yang et al. 2020) is the first work that applies knowledge distillation to GCNs by utilizing the local structure preserving module. And it can be seen as the one-teacher version of our proposed method. Thus, some of our parameter settings of the teachers and the student are very similar to LSP.
- Jing et.al. (Jing et al. 2021) propose a GNN-based knowledge amalgamation approach which is accomplished through a slimmable graph convolutional operation to learn varying-dimension features from teachers, together with a topological attribution map (TAM) scheme for learning the teachers' topological semantics.
- $L * l$. $L * l$ is a variant of our model which is composed of L l-layer GNN models. By comparing with this multi-teacher KD variant, we can further illustrate the superiority of our assemble strategy. For example, we use 1-layer, 2-layer and 3-layer GATs in PPI to form our model. Thus, $L * l$ consists of $3 * 1$, $3 * 2$ and $3 * 3$.

Experimental Settings

For a fair comparison, we utilize the experimental settings in (Yang et al. 2020) for all the methods. GAT and GCN are chosen as the GNN models. In teacher GAT, each hidden layer has 4 attention heads and 256 hidden features, and the output layer has 6 attention heads and K hidden features. In student GAT, there are 5 layers, each hidden layer has 2 attention heads and 68 hidden features, and the output layer has 2 attention heads and K hidden features. The settings of the number of hidden features in each layer are the same in GCN.

In all the methods, the optimizer is Adam, the learning rate is set to 0.005, training epochs are 500 and weight decay equals 0. We tune all other hyperparameters to the best results on the validation set. λ in the Equation (8) is set to 7, 3, 3 and 4 in four datasets.

GNN Type	Model	PPI / Ratio	CiteSeer / Ratio	PubMed / Ratio
GAT	AT	95.40 / 5.2	93.52 / 52.7	88.24 / 3.6
	FitNet	95.60 / 5.2	93.66 / 52.7	88.38 / 3.6
	LSP	95.96 / 5.2	94.03 / 52.7	88.75 / 3.6
	Jing et.al	96.83 / 3.2	94.65 / 10.2	89.92 / 10.2
	$L * 1$	96.73 / 4.1	94.42 / 19.9	89.48 / 7.7
	$L * 2$	96.62 / 1.7	94.50 / 17.6	89.07 / 1.6
	$L * 3$	96.52 / 1.1	93.78 / 15.7	89.64 / 0.9
	$L * 4$	-	-	89.28 / 0.6
	Optimal Teacher	97.38	95.78	83.76
	MSKD	**97.31** / 1.7	**95.47** / 17.6	**91.81** / 1.1
GCN	AT	54.51 / 16.7	87.17 / 24.5	82.96 / 21.1
	FitNet	54.70 / 16.7	87.45 / 24.5	83.11 / 21.1
	LSP	55.01 / 16.7	87.88 / 24.5	83.40 / 21.1
	Jing et.al	61.79 / 14.8	88.84 / 9.4	84.03 / 9.4
	$L * 1$	58.74 / 18.8	88.87 / 8.8	84.13 / 10.3
	$L * 2$	60.07 / 6.8	88.65 / 8.2	83.47 / 7.0
	$L * 3$	61.50 / 4.2	87.44 / 7.7	83.95 / 5.1
	$L * 4$	55.18 / 3.1	-	-
	Optimal Teacher	68.55	92.60	86.04
	MSKD	**63.67** / 5.3	**89.67** / 8.2	**84.89** / 6.9

Table 1: The F1-scores and the parameter ratios of the student to the teacher(s) in all the models.

Mechanism	PPI	CiteSeer	PubMed
Equal	97.24	95.30	91.67
Attention	97.31	95.47	91.81

Table 2: The F1-scores of our method with equal and attention allocations.

Model	F1-score	Params	Degree
1-layer GAT	84.27	1.54M	27.44
2-layer GAT	97.38	3.64M	30.21
3-layer GAT	97.29	5.74M	32.13
4-layer GAT	62.83	7.84M	-
PPI	-	-	29.32

Table 3: The F1-scores and the number of parameters of GATs with different number of layers, and the average node degree of PPI and the correctly classified node degrees by GATs with different number of layers.

Model	F1-score	Params of teachers
1-layer	95.55	1.54M
2-layer	95.96	3.64M
3-layer	95.42	5.74M
1+2-layer	96.96	5.19M
2+3-layer	96.92	9.39M
1+3-layer	96.35	7.29M
1+2+3-layer	**97.24**	10.93M
1+2+3+4-layer	96.70	18.77M

Table 4: The F1-scores of MSKD which is trained by GATs teachers with different number of layers, and the number of parameters of corresponding teachers in PPI.

Results

Here, we first experimentally verify that GNN models with different number of layers focus on the different node degrees in Table 3 to further explore the relationship of node degrees and multiple l-layer GNN teacher models, where 4-layer GAT has already degraded. In Table 3, on the one hand, the GNN models with more layers focus on the nodes with higher degrees which corresponds to the dense and complex local structures. On the other hand, the GNN models with less layers focus on the nodes with lower degrees which corresponds to the sparse and simple local structures. Thus, it is meaningful to comprehensively learn from these GNN models with different number of layers to achieve multi-scale degree-related knowledge distillation.

Table 1 gives the F1-scores (%) and the parameter ratios (%) of the student to the teacher(s) in all the methods. We have mentioned that the student models in all the methods have the same setup, thus, Ratio in Table 1 reflects the model compressive ability because of the different teacher models sizes which are denominators of Ratio. Notice that the GNN model only contains one hidden layer before model degradation on Cora by our proposed method, so we infer that a single teacher is enough based on the analysis before. And the experiment results also show that, i.e., the performances of one teacher and multi-teacher are almost the same. Thus, we do not compare performances on Cora.

According to Table 1, MSKD achieves higher F1-scores than state-of-the-art knowledge distillation methods on all three datasets. Notably, some well-learned compact students can even surpass the optimal teacher, which further demonstrates the superiority of MSKD. The parameter ratios of the student to the teacher(s) in all the models with GCN are also intuitively shown in Figure 3. Except for the models with GAT on CiteSeer, the lowest ratios prove that we achieve the best model compression efficiency over all baselines.

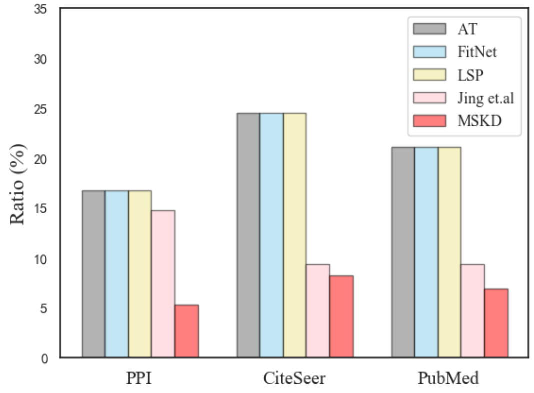


Figure 3: The parameter ratios of the student to the teacher(s) in all the models with GCN.

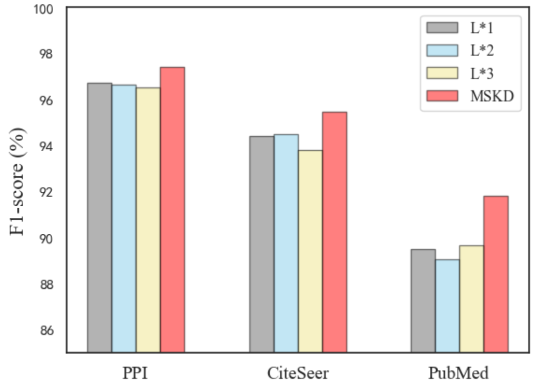


Figure 4: The F1-scores of MSKD and variants with the same number of layers.

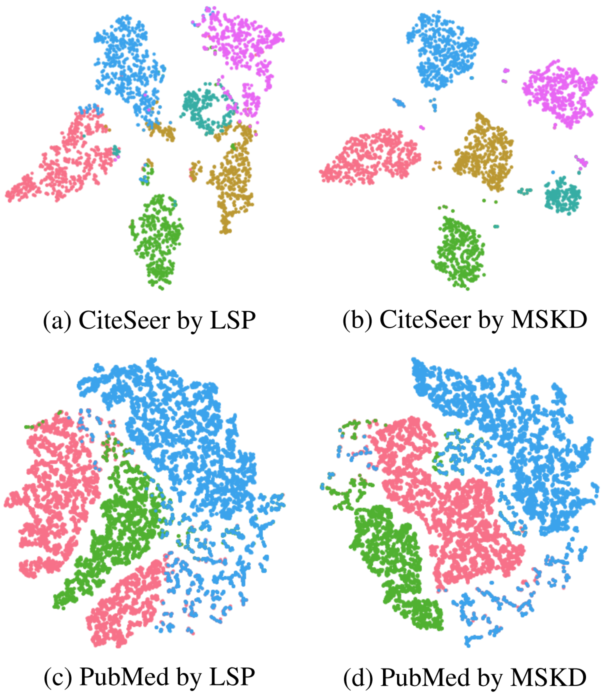
Ablation Study

To comprehensively evaluate our method, we provide ablation studies include the influences of the attention mechanism, the teacher GNN models with different number of layers, and the teacher selective strategy.

(1) In order to validate the effectiveness of the attention allocation, equal weight assignment is applied instead. In Table 2, the equal weight assignment causes lower F1-scores.

(2) Rather than learning from GNN models with the same number of layers, the student comprehensively learns multi-scale degree-related knowledge from GNN models with different number of layers. The performance differences are shown in Table 1 and intuitively shown in Figure 4, which demonstrates that.

(3) We select GNN models before model degradation as multiple teachers. Here, we show the other teacher selective results (without attention mechanism) in Table 4, where 1+2-layer means the teachers consist of 1-layer and 2-layer GATs, and others are the same. Table 4 demonstrate the effectiveness of our teacher selective strategy.


(a) CiteSeer by LSP

(b) CiteSeer by MSKD

(c) PubMed by LSP

(d) PubMed by MSKD

Figure 5: The visualizations of classified results on CiteSeer and PubMed by LSP and MSKD.

Visualization

We provide here visualizations of the GAT-version LSP and MSKD on CiteSeer and PubMed by the t-SNE (Bonin-Font, Ortiz, and Oliver 2008; Tang et al. 2016), which reduces the dimension of embedding to 2 and helps visualize nodes in a 2-dimensional space. The results are shown in Figure 5, where different colors represent different categories. On CiteSeer, comparing to LSP, MSKD's nodes with the same color are more aggregated, and the nodes with the different colors are more separated. On PubMed, the red nodes are wrong separated into two piles by LSP, but right separated by our method. These results demonstrate the superiority of MSKD.

Conclusion

The existing single-teacher knowledge distillation approaches which are based on a single GNN model, can only learn a student which contains the knowledge from the fixed scope of node degrees. To comprehensively and effectively learn the multi-scale degree-related knowledge, we proposed a multi-scale knowledge distillation method via attention allocation which selectively transfers the topological semantic knowledge from multiple l-layer GNN teacher models to the student. Experiments on single- and multi-label node classification tasks show that MSKD outperformed the state-of-the-art knowledge distillation methods, and even the teacher models to some extend. In addition, our model also achieved a significant model compression efficiency which is a key goal of knowledge distillation.

Acknowledgments

This research is supported by the National Natural Science Foundation of China under the grant No. 61976119 and the Natural Science Foundation of Tianjin under the grant No. 18ZXZNGX00310.

References

Bonin-Font, F.; Ortiz, A.; and Oliver, G. 2008. Visual navigation for mobile robots: A survey. *Journal of intelligent and robotic systems*, 53(3): 263–296.

Chen, D.; Mei, J.-P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; and Chen, C. 2021. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7028–7036.

Courbariaux, M.; Bengio, Y.; and David, J.-P. 2015. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, 3123–3131.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1025–1035.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Amalgamating Knowledge From Heterogeneous Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15709–15718.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proc. of ICLR*.

Lan, X.; Zhu, X.; and Gong, S. 2018. Knowledge distillation by on-the-fly native ensemble. *arXiv preprint arXiv:1806.04606*.

Lei, M.; Quan, P.; Ma, R.; Shi, Y.; and Niu, L. 2020. DigGCN: Learning Compact Graph Convolutional Networks via Diffusion Aggregation. *IEEE Transactions on Cybernetics*.

Luo, S.; Wang, X.; Fang, G.; Hu, Y.; Tao, D.; and Song, M. 2019. Knowledge amalgamation from heterogeneous networks by common feature learning. *arXiv preprint arXiv:1906.10546*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Sau, B. B.; and Balasubramanian, V. N. 2016. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*.

Sen, P.; Namata, G.; Bilgic, M.; Getoor, L.; Galligher, B.; and Eliassi-Rad, T. 2008. Collective classification in network data. *AI magazine*, 29(3): 93–93.

Shen, C.; Wang, X.; Song, J.; Sun, L.; and Song, M. 2019. Amalgamating knowledge towards comprehensive classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3068–3075.

Tang, J.; Liu, J.; Zhang, M.; and Mei, Q. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, 287–297.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph Attention Networks. *CoRR*, abs/1710.10903.

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341*.

Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; and Cheng, J. 2016. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4820–4828.

Xu, K.; Hu, W.; Leskovec, J.; and Jegelka, S. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Yang, Y.; Qiu, J.; Song, M.; Tao, D.; and Wang, X. 2020. Distilling knowledge from graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7074–7083.

Yao, H.; Zhang, C.; Wei, Y.; Jiang, M.; Wang, S.; Huang, J.; Chawla, N.; and Li, Z. 2020. Graph few-shot learning via knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6656–6663.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294.

Yu, X.; Liu, T.; Wang, X.; and Tao, D. 2017. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7370–7379.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhai, S.; Cheng, Y.; Lu, W.; and Zhang, Z. 2016. Doubly convolutional neural networks. *arXiv preprint arXiv:1610.09716*.

Zhang, W.; Sheng, Z.; Jiang, Y.; Xia, Y.; Gao, J.; Yang, Z.; and Cui, B. 2021. Evaluating deep graph neural networks. *arXiv preprint arXiv:2108.00955*.

Zhou, S.; Wang, Y.; Chen, D.; Chen, J.; Wang, X.; Wang, C.; and Bu, J. 2021. Distilling holistic knowledge with graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10387–10396.

Zitnik, M.; and Leskovec, J. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14): i190–i198.