

# Hierarchical Graph Convolutional Networks for Semi-supervised Node Classification

Fenyu Hu<sup>1,2\*</sup>, Yanqiao Zhu<sup>1,2\*</sup>, Shu Wu<sup>1,2†</sup>, Liang Wang<sup>1,2</sup> and Tieniu Tan<sup>1,2</sup>

<sup>1</sup> University of Chinese Academy of Sciences

<sup>2</sup> Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences  
{fenyu.hu, yanqiao.zhu}@cripac.ia.ac.cn, {shu.wu, wangliang, tnt}@nlpr.ia.ac.cn

## Abstract

Graph convolutional networks (GCNs) have been successfully applied in node classification tasks of network mining. However, most of these models based on neighborhood aggregation are usually shallow and lack the “graph pooling” mechanism, which prevents the model from obtaining adequate global information. In order to increase the receptive field, we propose a novel deep Hierarchical Graph Convolutional Network (H-GCN) for semi-supervised node classification. H-GCN first repeatedly aggregates structurally similar nodes to hyper-nodes and then refines the coarsened graph to the original to restore the representation for each node. Instead of merely aggregating one- or two-hop neighborhood information, the proposed coarsening procedure enlarges the receptive field for each node, hence more global information can be captured. The proposed H-GCN model shows strong empirical performance on various public benchmark graph datasets, outperforming state-of-the-art methods and acquiring up to 5.9% performance improvement in terms of accuracy. In addition, when only a few labeled samples are provided, our model gains substantial improvements.

## 1 Introduction

Graphs nowadays become ubiquitous owing to the ability to model complex systems such as social relationships, biological molecules, and publication citations. The problem of classifying graph-structured data is fundamental in many areas. Besides, since there is a tremendous amount of unlabeled data in nature and labeling data is often expensive and time-consuming, it is often challenging and crucial to analyze graphs in a semi-supervised manner. For instance, for semi-supervised node classification in citation networks, where nodes denote articles and edges represent citation, the task is to predict the label of every article with only a few labeled data.

As an efficient and effective approach to graph analysis, network embedding has attracted a lot of research interests. It aims to learn low-dimensional representations for nodes whilst still preserving the topological structure and node feature attributes. Many methods have been proposed for network embedding, which can be used in the node classification task, such as DeepWalk [Perozzi *et al.*, 2014] and node2vec [Grover and Leskovec, 2016]. They convert the graph structure into sequences by performing random walks on the graph. Then, the proximity between the nodes can be captured based on the co-occurrence statistics in these sequences. But they are unsupervised algorithms and cannot perform node classification tasks in an end-to-end fashion. Unlike previous random-walk-based approaches, employing neural networks on graphs has been studied extensively in recent years. Using an information diffusion mechanism, the graph neural network (GNN) model updates states of the nodes and propagate them until a stable equilibrium [Scarselli *et al.*, 2009]. Both of the highly non-linear topological structure and node attributes are fed into the GNN model to obtain the graph embedding. Recently, there is an increasing research interest in applying convolutional operations on the graph. These graph convolutional networks (GCNs) [Kipf and Welling, 2017; Veličković *et al.*, 2018] are based on the neighborhood aggregation scheme which generates node embedding by combining information from neighborhoods. Comparing with conventional methods, GCNs achieve promising performance in various graph analytical tasks such as node classification and graph classification [Defferrard *et al.*, 2016] and has shown effective for many application domains, for instance, recommendation [Wu *et al.*, 2019; Cui *et al.*, 2019], traffic forecasting [Yu *et al.*, 2018], and action recognition [Yan *et al.*, 2018].

Nevertheless, GCN-based models are usually shallow and lack the “graph pooling” mechanism, which restricts the scale of the receptive field. For example, there are only two layers in GCN [Kipf and Welling, 2017]. As each graph convolutional layer acts as the approximation of aggregation on the first-order neighbors, the two-layer GCN model only aggregates information from two-hop neighborhoods for each node. Because of the restricted receptive field, the model has difficulty in obtaining adequate global information. However, it has been observed from the reported results [Kipf and Welling, 2017] that simply adding more layers will de-

\*The first two authors contributed equally to this work.

†To whom correspondence should be addressed.

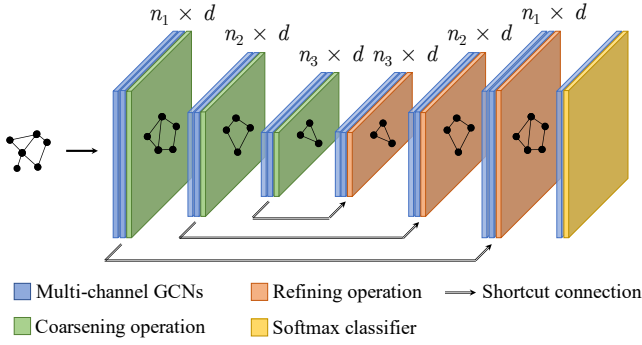


Figure 1: The workflow of H-GCN. In this illustration, there are seven layers with three coarsening layers, three symmetric refining layers, and one output layer. Coarsening layer at level  $i$  produces graph  $\mathcal{G}_{i+1}$  of  $n_{i+1}$  hyper-nodes with  $d$ -dimensional latent representations, vice versa for refining layers.

grade the performance. As explained in [Li *et al.*, 2018], each GCN layer acts as a form of Laplacian smoothing in essence, which makes the features of nodes in the same connected component similar. Thereby, adding too many convolutional layers will result in the output features over-smoothed and make them indistinguishable. Meanwhile, deeper neural networks with more parameters are harder to train. Although some recent methods [Chen *et al.*, 2018; Xu *et al.*, 2018; Ying *et al.*, 2018] try to get the global information through deeper models, they are either unsupervised models or need many training examples. As a result, they are still not capable of solving the semi-supervised node classification task directly.

To this end, we propose a novel architecture of Hierarchical Graph Convolutional Networks, H-GCN for brevity, for node classification on graphs<sup>1</sup>. Inspired from the flourish of applying deep architectures and the pooling mechanism into image classification tasks, we design a deep hierarchical model with coarsening mechanisms. The H-GCN model increases the receptive field of graph convolutions and can better capture global information. As illustrated in Figure 1, H-GCN mainly consists of several coarsening layers and refining layers. For each coarsening layer, the graph convolutional operation is first conducted to learn node representations. Then, a coarsening operation is performed to aggregate structurally similar nodes into hyper-nodes. After the coarsening operation, each hyper-node represents a local structure of the original graph, which can facilitate exploiting global structures on the graph. Following coarsening layers, we apply symmetric graph refining layers to restore the original graph structure for node classification tasks. Such a hierarchical model manages to comprehensively capture nodes’ information from local to global perspectives, leading to better node representations.

The main contributions of this paper are twofold. Firstly, to the best of our knowledge, it is the first work to design a deep hierarchical model for the semi-supervised node classification task. Compared to previous work, the proposed model

consists of more layers with larger receptive fields, which is able to obtain more global information through the coarsening and refining procedures. Secondly, we conduct extensive experiments on a variety of public datasets and show that the proposed method constantly outperforms other state-of-the-art approaches. Notably, our model gains a considerable improvement over other approaches with very few labeled samples provided for each class.

## 2 Related Work

In this section, we review some previous work on graph convolutional networks for semi-supervised node classification, hierarchical representation learning on graphs, and graph reduction algorithms.

### Graph Convolutional Networks

In the past few years, there has been a surge of applying convolutions on graphs. These approaches are essentially based on the neighborhood aggregation scheme and can be further divided into two branches: spectral approaches and spatial approaches.

The spectral approaches are based on the spectral graph theory to define parameterized filters. Bruna *et al.* (2014) first define the convolutional operation in the Fourier domain. However, its heavy computational burden limits the application to large-scale graphs. In order to improve efficiency, Defferrard *et al.* (2016) propose ChebNet to approximate the  $K$ -polynomial filters by means of a Chebyshev expansion of the graph Laplacian. Kipf and Welling (2017) further simplify the ChebNet by truncating the Chebyshev polynomial to the first-order neighborhood. DGCN [Zhuang and Ma, 2018] uses random walks to construct a positive mutual information matrix. Then, it utilizes that matrix along with the graph’s adjacency matrix to encode both local consistency and global consistency.

The spatial approaches generate node embedding by combining the neighborhood information in the vertex domain. MoNet [Monti *et al.*, 2017] and SplineCNN [Fey *et al.*, 2018] integrate the local signals by designing a universe patch operator. To generalize to unseen nodes in an inductive setting, GraphSAGE [Hamilton *et al.*, 2017] samples a fixed number of neighbors and employs several aggregation functions, such as concatenation, max-pooling, and LSTM aggregator. GAT [Veličković *et al.*, 2018] introduces the attention mechanism to model different influences of neighbors with learnable parameters. Gao *et al.* (2018) select a fixed number of neighborhood nodes for each feature and enables the use of regular convolutional operations on Euclidean spaces. However, the above two branches of GCNs are usually shallow and cannot obtain adequate global information as a consequence.

### Hierarchical Representation Learning on Graphs

Some work has been proposed for learning hierarchical information on graphs. Chen *et al.* (2018) and Liang *et al.* (2018) use a coarsening procedure to construct a coarsened graph of smaller size and then employ unsupervised methods, such as DeepWalk [Perozzi *et al.*, 2014] and node2vec [Grover and Leskovec, 2016] to learn node embedding based

<sup>1</sup>To make our results reproducible, all relevant source codes are publicly available at <https://github.com/CRIPAC-DIG/H-GCN>.

on that coarsened graph. Then, they conduct a refining procedure to get the original graph embedding. However, their two-stage methods are not capable of utilizing node attribute information and can neither perform node classification task in an end-to-end fashion. JK-Nets [Xu *et al.*, 2018] proposes general layer aggregation mechanisms to combine the output representation in every GCN layer. However, it can only propagate information across edges of the graph and are unable to aggregate information hierarchically. Therefore, the hierarchical structure of the graph cannot be learned by JK-Nets. To solve this problem, DiffPool [Ying *et al.*, 2018] proposes a pooling layer for graph embedding to reduce the size by a differentiable network. As DiffPool is designed for graph classification tasks, it cannot generate embedding for every node in the graph; hence it cannot be directly applied in node classification scenarios.

### Graph Reduction

Many approaches have been proposed to reduce the graph size without losing too much information, which facilitate downstream network analysis tasks such as community discovery and data summarization. There are two main classes of methods that reduce the graph size: graph sampling and graph coarsening. The first category is based on graph sampling strategy [Papagelis *et al.*, 2013; Hu and Lau, 2013; Chen *et al.*, 2017], which might lose key information during the sampling process. The second category applies graph coarsening strategies that collapse structure-similar nodes into hyper-nodes to generate a series of increasingly coarser graphs. The coarsening operation typically consists of two steps, i.e. grouping and collapsing. At first, every node is assigned to groups in a heuristic manner. Here a group refers to a set of nodes that constitute a hyper-node. Then, these groups are used to generate a coarser graph. For an unmatched node, Hendrickson and Leland (1995) randomly select one of its un-matched neighbors and merge these two nodes. Karypis and Kumar (1998) merge the two un-matched nodes by selecting those with the maximum weight edge. LaSalle and Karypis (2015) use a secondary jump during grouping.

However, these graph reduction approaches are usually used in unsupervised scenarios, such as community detection and graph partition. For semi-supervised node classification tasks, existing graph reduction methods cannot be used directly, as they are not capable of learning complex attributive and structural features of graphs. In this paper, H-GCN conducts graph reduction for non-Euclidean geometry like the pooling mechanism for Euclidean data. In this sense, our work bridges graph reduction for unsupervised tasks to the practical but more challenging semi-supervised node classification problems.

## 3 The Proposed Method

### 3.1 Preliminaries

#### Notations and Problem Definition

For the input undirected graph  $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ , where  $\mathcal{V}_1$  and  $\mathcal{E}_1$  are respectively the set of  $n_1$  nodes and  $e_1$  edges, let  $A_1 \in \mathbb{R}^{n_1 \times n_1}$  be the adjacency matrix describing its edge

weights and  $X \in \mathbb{R}^{n_1 \times d_1}$  be the node feature matrix, where  $d_1$  is the dimension of the attributive features. We use edge weights to indicate connection strengths between nodes. For the H-GCN network, the graph fed into the  $i^{\text{th}}$  layer is represented as  $\mathcal{G}_i$  with  $n_i$  nodes. The adjacency matrix and hidden representation matrix of  $\mathcal{G}_i$  are represented by  $A_i \in \mathbb{R}^{n_i \times n_i}$  and  $H_i \in \mathbb{R}^{n_i \times d_i}$  respectively.

Since coarsening layers and refining layers are symmetrical,  $A_i$  is identical to  $A_{l-i+1}$ , where  $l$  is the total number of layers in the network. For example, in the seven-layer model illustrated in Figure 1,  $\mathcal{G}_3$  is the input graph for the third layer and  $\mathcal{G}_5$  is the resulting graph from the fourth layer. After one coarsening operation and one refining operation,  $\mathcal{G}_3$  and  $\mathcal{G}_5$  share exactly the same topological structure  $A_3$ . As nodes will be assigned as a hyper-node, we define *node weight* as the number of nodes contained in a hyper-node.

Given the labeled node set  $\mathcal{V}_L$  containing  $m \ll n_1$  nodes, where each node  $v_i \in \mathcal{V}_L$  is associated with a label  $y_i \in \mathcal{Y}$ , our objective is to predict labels of  $\mathcal{V} \setminus \mathcal{V}_L$ .

### Graph Convolutional Networks

Graph convolutional networks achieve promising generalization in various tasks and our work is built upon the GCN module. At layer  $i$ , taking graph adjacency matrix  $A_i$  and hidden representation matrix  $H_i$  as input, each GCN module outputs a hidden representation matrix  $G_i \in \mathbb{R}^{n_i \times d_{i+1}}$ , which is described as:

$$G_i = \text{ReLU} \left( \tilde{D}_i^{-\frac{1}{2}} \tilde{A}_i \tilde{D}_i^{-\frac{1}{2}} H_i \theta_i \right), \quad (1)$$

where  $H_1 = X$ ,  $\text{ReLU}(x) = \max(0, x)$ , adjacency matrix with self-loop  $\tilde{A}_i = A_i + I$ ,  $\tilde{D}_i$  is the degree matrix of  $\tilde{A}_i$ , and  $\theta_i \in \mathbb{R}^{d_i \times d_{i+1}}$  is a trainable weight matrix. For ease of parameter tuning, we set output dimension  $d_i = d$  for all coarsening and refining layers throughout this paper.

### 3.2 The Overall Architecture

For a H-GCN network of  $l$  layers, the  $i^{\text{th}}$  graph coarsening layer first conducts a graph convolutional operation as formulated in Eq. (1) and then aggregates structurally similar nodes into hyper-nodes, producing a coarser graph  $\mathcal{G}_{i+1}$  and node embedding matrix  $H_{i+1}$  with fewer nodes. The corresponding adjacent matrix  $A_{i+1}$  and  $H_{i+1}$  will be fed into the  $(i+1)^{\text{th}}$  layer. Symmetrically, the graph refining layer also performs a graph convolution at first and then refines the coarsened graph to restore the finer graph structure. In order to boost optimization in deeper networks, we add shortcut connections [He *et al.*, 2016] across each coarsened graph and its corresponding refined part.

Since the topological structure of the graph changes between layers, we further introduce a node weight embedding matrix  $S_i$ , which transforms the number of nodes contained in each hyper-node into real-valued vectors. Both of the node weight embedding and  $H_i$  will be fed into the  $i^{\text{th}}$  layer. Besides, we add multiple channels by employing different GCNs to explore different feature subspaces.

The graph coarsening layers and refining layers altogether integrate different levels of node features and thus avoid over-smoothing during repeated neighborhood aggregation. After the refining process, we obtain a node embedding matrix

$H_{l-1} \in \mathbb{R}^{n_1 \times d}$ , where each row represents a node representation vector. In order to classify each node, we apply an additional GCN module followed by a softmax classifier on  $H_{l-1}$ .

### 3.3 The Graph Coarsening Layer

Every graph coarsening layer consists of two steps, i.e. graph convolution and graph coarsening. A GCN module is firstly used to extract structural and attributive features by aggregating neighborhoods' information as described in Eq. (1). For the graph coarsening procedure, we design the following two hybrid grouping strategies to assign nodes with similar structures into a hyper-node in the coarser graph. We first conduct structural equivalence grouping, followed by structural similarity grouping.

**Structural equivalence grouping (SEG).** If two nodes share the same set of neighbors, they are considered to be structurally equivalent. We then assign these two nodes to be a hyper-node. For example, as illustrated in Figure 2, nodes  $B$  and  $D$  are structurally equivalent, so these two nodes are allocated as a hyper-node. We mark all these structurally equivalent nodes and leave other nodes unmarked to avoid repetitive grouping operation on nodes.

**Structural similarity grouping (SSG).** Then, we calculate the *structural similarity* between the unmarked node pairs  $(v_j, v_k)$  as the normalized connection strength  $s(v_j, v_k)$ :

$$s(v_j, v_k) = \frac{A_{jk}}{\sqrt{D(v_j) \cdot D(v_k)}}, \quad (2)$$

where  $A_{jk}$  is the edge weight between  $v_j$  and  $v_k$ , and  $D(\cdot)$  is the node weight.

We iteratively take out an unmarked node  $v_j$  and calculate normalized connection strengths with all its unmarked neighbors. After that, we select its neighbor node  $v_k$  which has the largest structural similarity to form a new hyper-node and mark the two nodes. Particularly, if one node is left unmarked and all of its immediate neighbors are marked, it will be marked as well and constitutes a hyper-node by itself. For example, in Figure 2, node pair  $(C, E)$  has the largest structural similarity, so they are grouped together to form a hyper-node. After that, since only node  $A$  remains unmarked, it constitutes a hyper-node by itself.

Please note that if we take out unmarked nodes in a different order, the resulting hyper-graph will be different. The later we take a node out, the less its neighbors will be left unmarked. So, for a node with fewer neighbors, it has fewer probabilities to be grouped when it is taken out late. Therefore, we take out the unmarked nodes in ascending order according to the number of neighbors.

Using the above two grouping strategies, we are able to acquire all the hyper-nodes. For one hyper-node  $v_i$ , its edge weight to  $v_j$  is the summation over edge weights of  $v_j$ 's neighbor nodes contained in  $v_i$ . The updated node weights and edge weights will be used in Eq. (2) in the next coarsening layer.

In order to help restore the coarsened graph to original graph, we preserve the grouping relationship between nodes and their corresponding hyper-nodes in a matrix  $M_i \in$

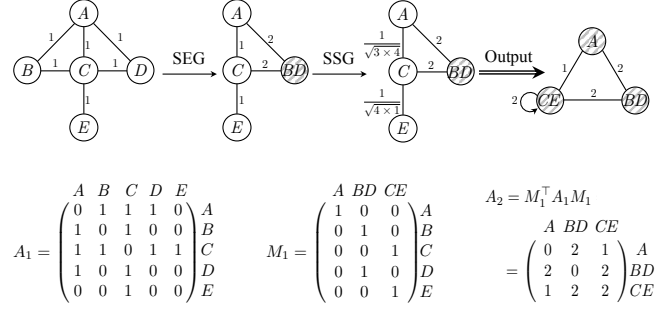


Figure 2: The graph coarsening operation of a toy graph. Numbers indicate edge weights and nodes in shadow are hyper-nodes. In SEG, node  $B$  and  $D$  share the same neighbors, so they are grouped into a hyper-node. In SSG, node  $C$  and  $E$  are grouped because they have the largest normalized connection weight. Node  $A$  constitutes a hyper-node by itself since it remains unmarked.

$\mathbb{R}^{n_i \times n_{i+1}}$ . Formally, at layer  $i$ , entry  $m_{jk}$  in the grouping matrix  $M_i$  is calculated as:

$$m_{jk} = \begin{cases} 1, & \text{if } v_j \text{ in } \mathcal{G}_i \text{ is grouped into } v_k \text{ in } \mathcal{G}_{i+1}; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

An example of the coarsening operation on a toy graph is given in Figure 2. Note that  $m_{11} = 1$  in this illustration, since node  $A$  constitutes its hyper-node by itself. Next, the hidden node embedding matrix is determined as:

$$H_{i+1} = M_i^T \cdot G_i. \quad (4)$$

In the end, we generate a coarser graph  $\mathcal{G}_{i+1}$ , whose adjacency matrix can be calculated as:

$$A_{i+1} = M_i^T \cdot A_i \cdot M_i. \quad (5)$$

The coarser graph  $\mathcal{G}_{i+1}$  along with the resulting representation matrix  $H_{i+1}$  will be fed into the next layer as input. The resulting node embedding to generate in each coarsening layer will then be of lower resolution. The graph coarsening procedure is summarized in Algorithm 1.

### 3.4 The Graph Refining Layer

To restore the original topological structure of the graph and further facilitate node classification, we stack the same numbers of graph refining layers as coarsening layers. Like the coarsening procedure, each refining layer contains two steps, namely generating node embedding vectors and restoring node representations.

To learn a hierarchical representation of nodes, a GCN is employed at first. Since we have saved the grouping relationship in the grouping matrix during the coarsening process, we utilize  $M_{l-i}$  to restore the refined node representation matrix of layer  $i$ . We further employ residual connections between the two corresponding coarsening and refining layers. In summary, node representations are computed by:

$$H_i = M_{l-i} \cdot G_i + G_{l-i}. \quad (6)$$

### 3.5 Node Weight Embedding and Multiple Channels

As depicted in Figure 2, since different hyper-nodes may have different node weights, we assume such consequent node

---

**Algorithm 1:** The graph coarsening operation

---

**Input:** Graph  $\mathcal{G}_i$  and node representation  $H_i$ **Output:** Coarsened graph  $\mathcal{G}_{i+1}$  and node representation  $H_{i+1}$ 

- 1 Calculate GCN output  $G_i$  according to Eq. (1)
  - 2 Initialize all nodes as unmarked  
/\* Structural equivalence grouping \*/
  - 3 Group and mark node pairs having the same neighbors  
/\* Structural similarity grouping \*/
  - 4 Sort all unmarked nodes in ascending order according to the number of neighbors
  - 5 **repeat**
  - 6   **for** each unmarked node  $v_j$  **do**
  - 7     **for** each unmark node  $v_k$  adjacent to  $v_j$  **do**
  - 8        Calculate  $s(v_j, v_k)$  according to Eq. (2)
  - 9     Group and mark the node pair  $(v_j, v_k)$  having the largest  $s(v_j, v_k)$
  - 10 **until** all nodes are marked
  - 11 Update node weights and edge weights
  - 12 Construct grouping matrix  $M_i$  according to Eq. (3)
  - 13 Calculate node representation  $H_{i+1}$  according to Eq. (4)
  - 14 Construct coarsened graph  $\mathcal{G}_{i+1}$  according to Eq. (5)
  - 15 **return**  $\mathcal{G}_{i+1}, H_{i+1}$
- 

weights could reflect the hierarchical characteristics of coarsened graphs. In order to better capture the hierarchical information, we use the node weight embedding to supplement information in  $H_i$ . Here we transform the node weight into real-valued vectors by looking up one randomly initialized node weight embedding matrix  $V \in \mathbb{R}^{|T| \times p}$ , where  $T$  is the set of node weights and  $p$  is the dimension of the embedding. We apply node weight embedding in every coarsening and refining layer. For graph  $\mathcal{G}_i$ , we obtain its node weight embedding  $S_i \in \mathbb{R}^{n_i \times p}$  by looking up  $V$  according to the node weight. For example, if one hyper-node contains three nodes, the third row of  $V$  will be selected as its node weight embedding. We then concatenate  $H_i$  and  $S_i$  and the resulting  $(d+p)$ -dimensional matrix will be fed into the next GCN layer subsequently.

Multi-channel mechanisms help explore features in different subspaces and H-GCN employs multiple channels on GCN to obtain rich information jointly at each layer. After obtained  $c$  channels  $[G_i^1, G_i^2, \dots, G_i^c]$ , we perform weighted average on these feature maps:

$$G_i = \sum_{j=1}^c w_j \cdot G_i^j, \quad (7)$$

where  $w_j$  is a trainable weight of channel  $j$ .

### 3.6 The Output Layer

Finally, in the output layer  $l$ , we use a GCN with a softmax classifier on  $H_{l-1}$  to output probabilities of nodes:

$$H_l = \text{softmax} \left( \text{ReLU} \left( \tilde{D}_l^{-\frac{1}{2}} \tilde{A}_l \tilde{D}_l^{-\frac{1}{2}} H_{l-1} \theta_l \right) \right), \quad (8)$$

where  $\theta_l \in \mathbb{R}^{d \times |\mathcal{Y}|}$  is a trainable weight matrix and  $H_l \in \mathbb{R}^{n_1 \times |\mathcal{Y}|}$  denotes the probabilities of nodes belonging to each class  $y \in \mathcal{Y}$ .

The loss function is defined as the cross-entropy of predictions over the labeled nodes:

$$\mathcal{L} = - \sum_{i=1}^m \sum_{y=1}^{|\mathcal{Y}|} \mathbb{I}(h_i = y_i) \log P(h_i, y_i), \quad (9)$$

where  $\mathbb{I}(\cdot)$  is the indicator function,  $y_i$  is the true label for  $v_i$ ,  $h_i$  is the prediction for labeled node  $v_i$ , and  $P(h_i, y_i)$  is the predicted probability that  $v_i$  is of class  $y_i$ .

### 3.7 Complexity Analysis and Model Comparison

In this section, we analyze the model complexity and compare it with mainstream graph convolutional models, such as GCN and GAT.

For GCN, preprocessing matrices  $\tilde{D}_i^{-\frac{1}{2}} \tilde{A}_i \tilde{D}_i^{-\frac{1}{2}}$  takes  $O(n^3)$ , and the training process for each layer takes  $O(|\mathcal{E}|CF)$ , where  $\mathcal{E}$  is the edge set and  $C, F$  are embedding dimensions. For GAT, the masked attention over all nodes requires  $O(n^2)$  in the training process.

For H-GCN, the preprocessing takes  $O(n \log n)$  to sort the unmarked nodes and  $O(mn)$  for SSG, where  $m$  is the average number of neighborhoods. For training, the complexity is also  $O(|\mathcal{E}|CF)$ . Therefore, H-GCN is as asymptotically efficient as GCN and is more efficient than GAT.

## 4 Experiments and Analysis

### 4.1 Experimental Settings

#### Datasets

For a comprehensive comparison with state-of-the-art methods, we use four widely-used datasets including three citation networks and one knowledge graph. We conduct semi-supervised node classification task in the transductive setting. The statistics of these datasets are summarized in Table 1. We set the node weight and edge weight of the graph to one for all four datasets. The dataset configuration follows the same setting in [Yang *et al.*, 2016; Kipf and Welling, 2017] for a fair comparison. For citation networks, documents and citations are treated as nodes and edges, respectively. For the knowledge graph, each triplet  $(e_1, r, e_2)$  will be assigned with separate relation nodes  $r_1$  and  $r_2$  as  $(e_1, r_1)$  and  $(e_2, r_2)$ , where  $e_1$  and  $e_2$  are entities and  $r$  is the relation between them. During training, only 20 labels per class are used for each citation network and only one label per class is used for NELL during training. Besides, 500 nodes in each dataset are selected randomly as the validation set. We do not use the labels of the validation set for model training.

#### Baseline Methods

To evaluate the performance of H-GCN, we compare our method with the following representative methods:

- **DeepWalk** [Perozzi *et al.*, 2014] generates the node embedding via random walks in an unsupervised manner, then nodes are classified by feeding the embedding vectors into an SVM classifier.



Dataset	Cora	Citeseer	Pubmed	NELL
Type	Citation network			Knowledge graph
# Vertices	2,708	3,327	19,717	65,755
# Edges	5,429	4,732	44,338	266,144
# Classes	7	6	3	210
# Features	1,433	3,703	500	5,414
Labeling rate	0.052	0.036	0.003	0.003

Table 1: Statistics of datasets used in experiments

Method	Cora	Citeseer	Pubmed	NELL
DeepWalk	67.2%	43.2%	65.3%	58.1%
Planetoid	75.7%	64.7%	77.2%	61.9%
GCN	81.5%	70.3%	79.0%	73.0%
GAT	83.0 $\pm$ 0.7%	72.5 $\pm$ 0.7%	79.0 $\pm$ 0.3%	—
DGCN	83.5%	72.6%	79.3%	74.2%
H-GCN	<b>84.5 <math>\pm</math> 0.5%</b>	<b>72.8 <math>\pm</math> 0.5%</b>	<b>79.8 <math>\pm</math> 0.4%</b>	<b>80.1 <math>\pm</math> 0.4%</b>

Table 2: Results of node classification in terms of accuracy

- **Planetoid** [Yang *et al.*, 2016] not only learns node embedding but also predicts the context in graph. It also leverages label information to build both transductive and inductive formulations.
- **GCN** [Kipf and Welling, 2017] produces node embedding vectors by truncating the Chebyshev polynomial to the first-order neighborhoods.
- **GAT** [Veličković *et al.*, 2018] generates node embedding vectors by modeling the differences between the node and its one-hop neighbors.
- **DGCN** [Zhuang and Ma, 2018] utilizes the graph adjacency matrix and the positive mutual information matrix to encode both local consistency and global consistency.

### Parameter Settings

We train our model using Adam optimizer with a learning rate of 0.03 for 250 epochs. The dropout is applied to all feature vectors with rates of 0.85. Besides, the  $\ell_2$  regularization factor is set to 0.0007. Considering different scales of datasets, we set the total number of layers  $l$  to 9 for citation networks and 11 for the knowledge graph, and apply four-channel GCNs in both coarsening and refining layers.

### 4.2 Node Classification Results

To demonstrate the overall performance of semi-supervised node classification, we compare the proposed method with other state-of-the-art methods. The performance in terms of accuracy is shown in Table 2. The best performance of each column is highlighted in boldface. The performance of our proposed method is reported based on the average of 20 measurements. Note that running GAT on the NELL dataset requires more than 64G memory; hence its performance on NELL is not reported.

The results show that the proposed method consistently outperforms other state-of-the-art methods, which verify the effectiveness of the proposed coarsening and refining mechanisms. Notably, compared with citation networks, H-GCN surpasses other baselines by larger margins on the NELL dataset. To be specific, the accuracy of H-GCN exceeds GCN

Method	20	15	10	5
GCN	79.0%	76.9%	72.2%	69.0%
GAT	79.0%	77.3%	75.4%	70.3%
DGCN	79.3%	77.4%	76.7%	70.1%
H-GCN	<b>79.8%</b>	<b>79.3%</b>	<b>78.6%</b>	<b>76.5%</b>

Table 3: Results of node classification in terms of accuracy on Pubmed with labeled vertices varying from 20 per class to 5.

and DGCN by 7.1% and 5.9% on NELL dataset respectively. We analyze the results as follows.

Regarding traditional random-walk-based algorithms such as DeepWalk and Planetoid, their performance is relatively poor. DeepWalk cannot model the attribute information, which heavily restricts its performance. Though Planetoid combines supervised information with an unsupervised loss, there is information loss of graph structure during random sampling. To avoid that problem, GCN and GAT employ the neighborhood aggregation scheme to boost performance. GAT outperforms GCN as it can model different relations to different neighbors rather than with a pre-defined order. DGCN further jointly models both local and global consistency, yet its global consistency is still obtained through random walks. As a result, the information in the graph structure might lose in DGCN as well. On the contrary, the proposed H-GCN manages to capture global information through different levels of convolutional layers and achieves the best results among all four datasets.

Besides, on the NELL dataset, there are fewer training samples per class than in citation networks. Under such circumstance, training nodes are further away from testing nodes on average. The baseline models with the restricted receptive field are unable to propagate the features and the supervised information of the training nodes to other nodes sufficiently. As a result, the proposed H-GCN with increased receptive fields and deeper layers obtains more promising improvements than baselines.

### 4.3 Impact of Scale of Training Data

We suppose that a larger receptive field in the convolutional model promotes the propagation of features and labels on graphs. To verify the proposed H-GCN can get a larger receptive field, we reduce the number of training samples to check if H-GCN still performs well when limited labeled data is given. As in nature, there are plenty of unlabeled data; it is also of great significance to train the model with limited labeled data. In this section, we conduct experiments with different numbers of labeled instances on the Pubmed dataset. We vary the number of labeled nodes from 20 to 5 per class, where the labeled data is randomly chosen from the original training set. All parameters are the same as previously described. The corresponding performance in terms of accuracy is reported in Table 3.

From the table, it can be observed that our method outperform other baselines in all cases. With the number of labeled data decreasing, our method obtains a more considerable margin over these baseline algorithms. Especially when only five labeled nodes per class ( $\approx 0.08\%$  labeling rate) are given,

Method	Cora	Citeseer	Pubmed	NELL
H-GCN without coarsening and refining layers	80.3%	70.5%	76.8%	75.9%
H-GCN without node weight embeddings	84.2%	72.4%	79.5%	79.6%
H-GCN	84.5%	72.8%	79.8%	80.1%

Table 4: Results of the ablation study

the accuracy of H-GCN exceeds GCN, DGCN, and GAT by 7.5%, 6.4%, and 6.2% respectively. When the number of training data decreases, it is more likely for an unlabeled node to be further away from these labeled nodes. Only when the receptive field is large enough can information from those training nodes be captured. As the receptive field of GCN and GAT does not exceed 2-hop neighborhoods, supervised information contained in the training nodes cannot propagate sufficiently to other nodes. Therefore, these baselines downgrade considerably. However, owing to its larger receptive field, the performance of H-GCN declines slightly when labeled data decreases dramatically. Overall, it is verified that the proposed H-GCN with increased receptive fields is well-suited when training data is extremely scarce and thereby is of significant practical values.

#### 4.4 Ablation Study

To verify the effectiveness of the proposed coarsening and refining layers, we conduct ablation study on coarsening and refining layers and node weight embeddings respectively in this section. The results are shown in Table 4.

**Coarsening and refining layers.** We remove all coarsening and refining operations of H-GCN and compare its performance with the original H-GCN. Different from simply adding too many GCN layers, we preserve the short-cut connection between the symmetric layers in the ablation study. From the results, it is evident that the proposed H-GCN has better performance compared to H-GCN without coarsening mechanisms on all datasets. It can be verified that the coarsening and refining mechanisms contribute to the performance improvements since they can obtain global information with larger receptive fields.

**Node weight embeddings.** To study the impact of node weight embeddings, we compare H-GCN with no node weight embeddings used. It can be seen from results that the model with node weight embeddings performs better, which verifies the necessity to add this embedding vector in the node embeddings.

#### 4.5 Sensitivity Analysis

Last, we analyze hyper-parameter sensitivity. Specifically, we investigate how different numbers of coarsening layers and different numbers of channels will affect the results respectively. The performance is reported in terms of accuracy on all four datasets. While one parameter studied in the sensitivity analysis is changed, other hyper-parameters remain the same.

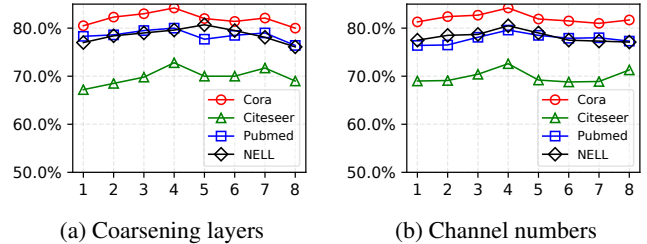


Figure 3: Results of H-GCN with varying layers and channels in terms of accuracy.

**Effects of coarsening layers.** Since the coarsening layers in our model control the granularity of the receptive field enlargement, we experiment with one to eight coarsening and symmetric refining layers, where the results are shown in Figure 3(a). It is seen that the performance of H-GCN achieves the best when there are four coarsening layers on three citation networks and five on the knowledge graph. It is suspected that, since less labeled nodes are supplied on NELL than others, deeper layers and larger receptive fields are needed. However, when adding too many coarsening layers, the performance drops due to overfitting.

**Effects of channel numbers.** Next, we investigate the impact of different amounts of channels on the performance. Multiple channels benefit the graph convolutional network model, since they explore different feature subspaces, as shown in Figure 3(b). From the figure, it can be found that the performance improves with the number of channels increasing until four channels, which demonstrates that more channels help capture accurate node features. Nevertheless, too many channels will inevitably introduce redundant parameters to the model, leading to overfitting as well.

## 5 Conclusion

In this paper, we have proposed a novel hierarchical graph convolutional networks for the semi-supervised node classification task. The H-GCN model consists of coarsening layers and symmetric refining layers. By grouping structurally similar nodes to hyper-nodes, our model can get a larger receptive field and enable sufficient information propagation. Compared with other previous work, our proposed H-GCN is deeper and can fully utilize both local and global information. Comprehensive experiments have confirmed that the proposed method consistently outperformed other state-of-the-art methods. In particular, our method has achieved substantial gains over them in the case that labeled data is extremely scarce.

## Acknowledgements

This work is jointly supported by National Natural Science Foundation of China (61772528) and National Key Research and Development Program (2016YFB1001000, 2018YFB1402600).

## References

- [Bruna *et al.*, 2014] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs. In *ICLR*, 2014.
- [Chen *et al.*, 2017] Haibo Chen, Jianfei Zhao, Xiaoji Chen, Ding Xiao, and Chuan Shi. Visual analysis of large heterogeneous network through interactive centrality based sampling. In *ICNSC*, pages 378–383, May 2017.
- [Chen *et al.*, 2018] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. Harp: Hierarchical representation learning for networks. In *AAAI*, 2018.
- [Cui *et al.*, 2019] Zeyu Cui, Zekun Li, Shu Wu, Xiao-Yu Zhang, and Liang Wang. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *WWW*, pages 307–317, 2019.
- [Defferrard *et al.*, 2016] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3844–3852, 2016.
- [Fey *et al.*, 2018] Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. SplineCNN: Fast geometric deep learning with continuous B-spline kernels. In *CVPR*, 2018.
- [Gao *et al.*, 2018] Hongyang Gao, Zhengyang Wang, and Shuiwang Ji. Large-scale learnable graph convolutional networks. In *KDD*, pages 1416–1424, 2018.
- [Grover and Leskovec, 2016] Aditya Grover and Jure Leskovec. Node2vec: Scalable feature learning for networks. In *KDD*, pages 855–864, 2016.
- [Hamilton *et al.*, 2017] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pages 1025–1035, 2017.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hendrickson and Leland, 1995] Bruce Hendrickson and Robert Leland. A multilevel algorithm for partitioning graphs. In *Supercomputing*, 1995.
- [Hu and Lau, 2013] Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *CoRR*, abs/1308.5865, 2013.
- [Karypis and Kumar, 1998] George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392, 1998.
- [Kipf and Welling, 2017] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.
- [LaSalle and Karypis, 2015] Dominique LaSalle and George Karypis. Multi-threaded modularity based graph clustering using the multilevel paradigm. *J. Parallel Distrib. Comput.*, 76(C):66–80, February 2015.
- [Li *et al.*, 2018] Qimai Li, Zhichao Han, and Xiao ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, pages 3538–3545, 2018.
- [Liang *et al.*, 2018] Jiongqian Liang, Saket Gurukar, and Srinivasan Parthasarathy. MILE: A Multi-Level Framework for Scalable Graph Embedding. *ArXiv e-prints*, 2018.
- [Monti *et al.*, 2017] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodolà, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *CVPR*, pages 5425–5434, 2017.
- [Papagelis *et al.*, 2013] Manos Papagelis, Gautam Das, and Nick Koudas. Sampling online social networks. *TKDE*, 25(3):662–676, March 2013.
- [Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, pages 701–710, 2014.
- [Scarselli *et al.*, 2009] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *TNN*, 20(1):61–80, 2009.
- [Veličković *et al.*, 2018] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.
- [Wu *et al.*, 2019] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *AAAI*, 2019.
- [Xu *et al.*, 2018] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. In *ICML*, pages 5453–5462, 2018.
- [Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018.
- [Yang *et al.*, 2016] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, pages 40–48, 2016.
- [Ying *et al.*, 2018] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, pages 4800–4810. 2018.
- [Yu *et al.*, 2018] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *IJCAI*, pages 3634–3640, 2018.
- [Zhuang and Ma, 2018] Chenyi Zhuang and Qiang Ma. Dual graph convolutional networks for graph-based semi-supervised classification. In *WWW*, pages 499–508, 2018.