

Graph Convolutional Neural Networks with Node Transition Probability-based Message Passing and DropNode Regularization

Tien Huu Do^{a,b,*}, Duc Minh Nguyen^{a,b}, Giannis Bekoulis^{a,b}, Adrian Munteanu^a, Nikos Deligiannis^{a,b}

^a *Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium*

^b *imec, Kapeldreef 75, B-3001 Leuven, Belgium*

Abstract

Graph convolutional neural networks (GCNNs) have received much attention recently, owing to their capability in handling graph-structured data. Among the existing GCNNs, many methods can be viewed as instances of a neural message passing motif; features of nodes are passed around their neighbors, aggregated and transformed to produce better nodes' representations. Nevertheless, these methods seldom use node transition probabilities, a measure that has been found useful in exploring graphs. Furthermore, when the transition probabilities are used, their transition direction is often improperly considered in the feature aggregation step, resulting in an inefficient weighting scheme. In addition, although a great number of GCNN models with increasing level of complexity have been introduced, the GCNNs often suffer from over-fitting when being trained on small graphs. Another issue of the GCNNs is over-smoothing, which tends to make nodes' representations indistinguishable. This work presents a new method to improve the message passing process based on node transition probabilities by properly considering the transition direction, leading to a better weighting scheme in nodes' features aggregation compared to the existing counterpart. Moreover, we propose a novel regularization method termed *DropNode* to address

*Corresponding author

Email addresses: thdo@etrovub.be (Tien Huu Do), mdnguyen@etrovub.be (Duc Minh Nguyen), gbekouli@etrovub.be (Giannis Bekoulis), acmuntea@etrovub.be (Adrian Munteanu), ndeligia@etrovub.be (Nikos Deligiannis)

the over-fitting and over-smoothing issues simultaneously. DropNode randomly discards part of a graph, thus it creates multiple deformed versions of the graph, leading to data augmentation regularization effect. Additionally, DropNode lessens the connectivity of the graph, mitigating the effect of over-smoothing in deep GCNNs. Extensive experiments on eight benchmark datasets for node and graph classification tasks demonstrate the effectiveness of the proposed methods in comparison with the state of the art.

Keywords: Graph convolutional neural networks, graph classification, node classification, geometric deep learning.

1. Introduction

In the recent years, we have witnessed an increasing interest in deep learning-based techniques for solving problems with data living in non-Euclidean domains, i.e., graphs and manifolds. In order to handle data in such domains, graph neural
5 networks (GNNs) have been proposed. There are several types of GNNs, however recently graph convolutional neural networks (GCNNs) have attracted a great deal of attention owing to their state-of-the-art performance in a number of tasks including text (Yao et al., 2019) and image (Quek et al., 2011) classification, semantic segmentation (Qi et al., 2017), question answering (Teney et al., 2017)
10 and matrix completion (Do et al., 2019). There are two main approaches for GCNNs, namely, *spectral* and *spatial* (Zhou et al., 2018) approaches. The former refers to models which define convolution operators based on graph Fourier transform (Defferrard et al., 2016; Henaff et al., 2015; Bruna et al., 2013; Ortega et al., 2018). On the other hand, models that follow the latter
15 approach directly generalize the classical convolution operator to non-Euclidean domains (i.e., by spatially aggregating information over nodes' neighborhoods in the graphs (Hamilton et al., 2017; Veličković et al., 2017; Niepert et al., 2016; Gao et al., 2018; Atwood & Towsley, 2016)). The spatial approach has become
20 scalability to large graphs.

In general, the operations of most GCNN models can be expressed in two phases, namely, the (i) *Message Passing* and (ii) *Readout* phases. In phase (i), a node in the graph receives messages from its neighbors, where each message is computed either from features of a neighboring node or features of an edge attached to the node. Two steps that are often performed in the message passing phase are the *Aggregate* and *Update* steps (Zhou et al., 2018). In the Aggregate step, a node gathers messages from its neighbors by performing the weighted sum operation (Kipf & Welling, 2016; Hamilton et al., 2017; Zhang et al., 2018) or employing a long short-term memory neural network (LSTM) (Hamilton et al., 2017). In the Update step, a non-linear transformation is then applied on the aggregated messages to produce the updated node representations. Common transformations used in the Update step include pre-defined functions, such as sigmoid and ReLU (Bruna et al., 2013; Defferrard et al., 2016), or learnable functions, such as gated recurrent units (GRU) (Li et al., 2015) and fully connected neural networks (Kearnes et al., 2016; Battaglia et al., 2016; Hamilton et al., 2017; Duvenaud et al., 2015; Kipf & Welling, 2016; Zhang et al., 2018). In GCNN models, the Message Passing phase is often implemented using *graph convolutional layers*. By stacking multiple graph convolutional layers, one can build a deep GCNN model and enable (a) passing messages between nodes that are more than one hop apart (i.e., nodes connected via other nodes) and (b) learning abstract intermediate node embeddings through a series of non-linear transformations. In phase (ii), the learned node embeddings are converted to the final representations for the nodes or for the whole graph. These representation are then fed to a classifier to produce suitable outputs depending on downstream tasks.

In GCNN models, the non-Euclidean characteristics of the data are captured mainly in the Message Passing phase. As such, designing an effective message passing scheme is of high importance, however it remains a challenging task. To address this challenge, various message passing formulations have been proposed (Gilmer et al., 2017; Duvenaud et al., 2015; Kipf & Welling, 2016; Hamilton et al., 2017). Recently, formulations based on node transition probabilities have

been proven beneficial for multiple graph-based classification tasks (Atwood
 & Towsley, 2016; Zhang et al., 2018). Nevertheless, such formulations have
 not received enough attention from the research community. In addition, other
 55 challenges such as the *over-fitting* and *over-smoothing* problems (e.g., especially
 when deploying *deep* GCNNs) often arise. To mitigate over-fitting, popular
 regularization methods such as ℓ_1 , ℓ_2 and *dropout* (Srivastava et al., 2014) are
 usually employed. Nevertheless, performance gains brought by these methods
 generally diminish as GCNNs become deeper (Kipf & Welling, 2016). It is worth
 60 noting that over-fitting is general, i.e., it may happens for different types of
 neural networks, including GCNNs. On the other hand, over-smoothing (Li
 et al., 2018; Chen et al., 2019) is specific for deep GCNNs as a result of the
 inherent smoothing effect of these models. Specifically, when many graph con-
 volutional layers are stacked together, the obtained representations of nodes in
 65 different classes (or clusters) become indistinguishable and inseparable, leading
 to degraded performance in downstream tasks such as node classification.

In this work, we aim to develop effective GCNN models to handle graph-
 structured data by addressing the aforementioned challenges. Firstly, we design
 a GCNN message passing scheme employing the transition-based approach.
 70 Specifically, we observe that the transition directions employed in existing
 works (Atwood & Towsley, 2016; Zhang et al., 2018) are sub-optimal since
 high weights (i.e., influence) are assigned to “popular” nodes (i.e., the ones
 with multiple connections). Based on this observation, we propose a novel
 message passing scheme that eliminates the aforementioned issue by assigning
 75 balanced weights to all the nodes in the graph. We experimentally validate the
 effectiveness of the proposed scheme in multiple graph classification tasks. In
 addition, we introduce a novel method called *DropNode* to address the over-fitting
 and over-smoothing problems on GCNNs. The idea behind this method is that
 the underlying structure of the graph is modified during training by randomly
 80 sub-sampling nodes from the graph. Despite its simplicity, the DropNode method
 is highly effective, especially in training deep GCNN models.

To summarize, our contributions in this paper are three-fold:

- We propose a novel neural message passing scheme for GCNNs, which leverages the node transition probabilities and also takes the “popularity” of nodes into account;
- We propose DropNode, a simple and effective regularization method which is applicable to different GCNN models;
- We carry out comprehensive experiments on eight real-world benchmark datasets on both node and graph classification tasks to evaluate the proposed models. Experimental results showcase that our models are able to obtain improved performance over state-of-the-art models.

The remainder of the paper is organized as follows. Section 2 reviews the related work on GCNNs and the commonly-used regularization methods. Section 3 introduces the notation and problem formulation, while Section 4 describes in detail the proposed method. We present the experimental results of the proposed models in Section 5 and we conclude our work in Section 6.

2. Related Work

2.1. Neural Message Passing

Most existing GCNN models perform one or multiple *message passing* steps, in which nodes’ or edges’ features are aggregated following specific message passing schemes (Gilmer et al., 2017). Early works, which follow the spectral approach, aggregate messages by summing over transformed features of the nodes where the transformations are parameterized by the eigenvectors of the Laplacian graph (Bruna et al., 2013; Defferrard et al., 2016). As these eigenvectors correspond to the Fourier bases of the graph, each message passing step can be seen as a convolution operation on the features of the graph. More recently, Kipf et al. (Kipf & Welling, 2016) proposed the Graph Convolutional Network (GCN) model which aggregates node features by employing the normalized graph adjacency matrix as a transformation matrix (in this case, the transformation is a linear projection). Since the message passing scheme of the GCN (in its

final form) does not rely on the Fourier analysis of the graph, GCN can be seen as a spatial-based model. Following GCN (Kipf & Welling, 2016), various models based on the spatial approach have been recently proposed. For instance, messages are aggregated by averaging nodes’ features in (Hamilton et al., 2017),
115 summing over concatenated features of nodes and edges in (Duvenaud et al., 2015), or by applying neural networks on nodes’ features (Hamilton et al., 2017) and edges’ features in (Battaglia et al., 2016; Schütt et al., 2017). Most of these spatial-based models utilize the connections between nodes in the graph directly as weights when aggregating node and edge features. The weight of the
120 connections can be pre-computed, or can be learned by leveraging the attention mechanisms, as in the attention graph neural networks (GAT) model (Veličković et al., 2017). In addition, other models such as the node transition-based ones employ transition probabilities between nodes, where these probabilities are calculated based on random walk procedures (Atwood & Towsley, 2016; Zhang
125 et al., 2018).

Our message passing scheme is inspired by the node transition-based models [see (Atwood & Towsley, 2016; Zhang et al., 2018)]. The key difference between the proposed method and these works is that we do not overestimate the influence of the popular nodes, which leads to the performance improvement on several
130 graph classification tasks.

2.2. Graph Regularization Methods

As the second contribution of this work is DropNode, a regularization technique for GCNNs, this section is dedicated to review existing regularization methods for GCNNs. Additionally, the relations between our method and the
135 existing methods are presented.

Dropout (Srivastava et al., 2014) is a commonly used regularization technique for various types of neural networks, including GCNNs. In dropout, a percentage of neural units is kept active during training using Bernoulli distribution, which in turn reduces the co-adapting effect of neural units. In testing, by using the
140 full architecture, the network is able to approximate the averaging of many

thinned models, leading to better performance. Unlike dropout, our method (DropNode) randomly removes nodes of the underlying graphs during the training of GCNNs. Conceptually, DropNode alters the data instead of neural units of network architecture. DropNode is similar to the work of Chen *et al.* (Chen et al., 2018) in terms of random node sampling. However, we take a further step by adding an *upsampling* operation, which recovers the dropped nodes from the previous sampling step. This arrangement allows us to yield the data augmentation effect following the U-Net principle (Ronneberger et al., 2015). In the model proposed by Gao *et al.* (Gao & Ji, 2019), pooling and unpooling operations are used to create a graph-based U-Net architecture. However, the pooling uses **top-k** selection based on scores of nodes; the scores are computed using a scalar projection of node features. This is different from our method as we randomly sample nodes using the Bernoulli distribution. As a result, the model in (Gao & Ji, 2019) does not have the regularization effect because the architecture is fixed across training and testing. In the work of (Rong et al., 2019), edges are randomly removed to regularize GCNNs, which is similar to our method in terms of modifying the underlying graph. However, as the number of nodes of the graph is kept unchanged, downsampling and upsampling operations are not used. As a result, while the method in (Rong et al., 2019) is easy to integrate to various GNN architectures, the regularization effect is not as strong as the proposed method, which alters both edges and nodes.

3. Preliminaries

3.1. Notation and Problem Formulation

Let $G = (V, E)$ denote an undirected graph, where V is the set of nodes and E is the set of edges. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix of G , where \mathbf{A}_{ij} is the weight of the corresponding edge connecting node i and node j . As G is an undirected graph, the adjacency matrix \mathbf{A} is symmetric (i.e., $\mathbf{A}_{ij} = \mathbf{A}_{ji}$). The diagonal matrix \mathbf{D} (where $\mathbf{D}_{ii} = \sum_{j=1}^N \mathbf{A}_{ij}$) denotes the degree matrix of G . The unnormalized Laplacian matrix of G is given by $\mathbf{L} = \mathbf{D} - \mathbf{A}$ (Ortega

et al., 2018). The graph G can be fully represented using either the Laplacian or the adjacency matrix. Each node of G is characterized by a feature vector $\mathbf{x} \in \mathbb{R}^F$, where F is the feature dimensionality. Thus, all the features of the nodes in G can be represented in the form of a matrix, $\mathbf{X} \in \mathbb{R}^{N \times F}$ where each row corresponds to a node and each column corresponds to a feature.

Two common tasks that graph neural networks often address are (i) node and (ii) graph classification. Task (i) concerns predicting a label for each node in a given graph. Specifically, for a graph $G = (V, E)$, the set of labelled nodes can be symbolized as V_L and the set of unlabelled nodes can be symbolized as V_U , such that $V = V_L \cup V_U$. The goal of task (i) is to predict the labels for the nodes in V_U . In the case that the features and the connections of these nodes are available during the training phase, the task follows the so-called *transductive setting* (Kipf & Welling, 2016; Veličković et al., 2017). On the other hand, in the case that neither the features nor the connections of nodes in V_U are given during training, the task follows the *inductive setting* (Hamilton et al., 2017; Chen et al., 2018). Different from the node classification task (i.e., task (i)), the task (ii) aims at predicting a label for each graph. For this task, only the inductive setting is considered, that is, given a training set of graphs $S_T = \{G_1, G_2, \dots, G_M\}$ and the corresponding set of labels $L_T = \{1, 2, \dots, C\}^M$ where C is the number of classes and M is the number of examples in the training set, the goal is to learn a prediction function that maps an input graph to a label in L_T . In this work, we will consider both the node and graph classification tasks.

3.2. Random Walk and Node Transition Matrix

Random Walk is one of the most effective methods to explore a graph (see Leskovec & Faloutsos (2006)). Given a graph $G = (V, E)$, a length constant l , and a starting node $u_0 \in V$, this method generates a “walk” of length l , represented by a sequence of nodes $\{u_0, u_1, \dots, u_l\}$. A node is sampled for step k in the walk, $k \in [1, l]$, following a distribution $P(u_k | u_{k-1})$, which is expressed

by:

$$P(u_k = v_j | u_{k-1} = v_i) = \begin{cases} P_{ij}, & \text{if } (v_i, v_j) \in E \\ 0, & \text{if } (v_i, v_j) \notin E. \end{cases} \quad (1)$$

In (1), P_{ij} is the probability of transitioning from node v_i to node v_j . As such, P_{ij} is often referred to as the *node transition probability*. The standard way to compute the transition probabilities is by dividing the weight of the edge (v_i, v_j) by the degree of node v_i (i.e., $P_{ij} = \frac{\mathbf{A}_{ij}}{\mathbf{D}_{ii}}$). In general, $P_{ij} \neq P_{ji}$ since the degree of node v_i is not equal to the degree of node v_j . All the transition probabilities of the nodes in G can be represented in the form of a matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$, where:

$$\mathbf{P} = \mathbf{D}^{-1} \mathbf{A}. \quad (2)$$

Note that $\sum_{j=1}^N \mathbf{P}_{ij} = 1, \forall i \in [1, N]$.

4. The Proposed Models

195 In this section, we present in detail the contribution of this work. At the beginning, we describe existing message passing schemes (Kipf & Welling, 2016; Zhang et al., 2018). Afterwards, we introduce the proposed scheme (which follows node transition-based approaches) that leads to a new graph convolutional layer formulation. Subsequently, we present DropNode, a simple yet effective generic
200 regularization method, and we compare it with existing methods that are often employed to regularize GCNN models.

4.1. Graph Convolutional Layers

4.1.1. Existing Formulations

As presented earlier in the Introduction section, the formulation of a graph convolutional layer in a GCNN model fully expresses the underlying message passing scheme. Various graph convolutional layers have been proposed in the literature that generally follow the same formulation:

$$\mathbf{H}^{(l+1)} = \sigma(\mathbf{M}\mathbf{H}^{(l)}\mathbf{W}), \quad (3)$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{N \times F}$ is the input to the l -th layer, $\mathbf{W} \in \mathbb{R}^{F \times K}$ is the weight matrix
of this layer, $\mathbf{M} \in \mathbb{R}^{N \times N}$ is a matrix containing the *aggregation coefficients*, and
 σ is a non-linear activation function such as the sigmoid or the ReLU.

The message passing scheme, as shown in (3), can be broken down into two
substeps, namely the *aggregate* and the *update* substeps. In the former, node
features are first transformed by a linear projection, which is parameterized by
learnable weights \mathbf{W} . Subsequently, for a node in the graph, the features of its
neighbors are aggregated by performing a weighted sum with the corresponding
weights defined in \mathbf{M} . This substep can be seen as a *message* aggregation stage
from the neighbors of each node. In the *update* substep, a non-linear transforma-
tion is applied to the aggregated features to produce new representations $\mathbf{H}^{(l+1)}$
for the nodes.

The aggregation coefficient matrix \mathbf{M} defines the message passing scheme,
i.e., the way in which features are exchanged between nodes. An entry \mathbf{M}_{ij} of
 \mathbf{M} represents the coefficient assigned to features of a source node j when being
aggregated toward a destination node i . As such, the i -th row of \mathbf{M} specifies the
weights assigned to features of all the nodes when being aggregated toward node
 i . The j -th column of \mathbf{M} specifies the weights assigned to a node j when being
aggregated toward all the other nodes. Intuitively, the rows and the columns
of matrix \mathbf{M} determine the influence of the nodes on a destination node, and
the influence of a source node on the other nodes, respectively. By defining the
aggregation coefficient matrix \mathbf{M} , one can specify a message passing scheme. For
instance, in a GCN model (Kipf & Welling, 2016), \mathbf{M} is computed by:

$$\mathbf{M} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}, \quad (4)$$

with $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_N$ and $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is an identity matrix. By adding the identity
matrix \mathbf{I}_N , self-connections are taken into account, i.e., a node aggregates features
from both its neighbors and itself. $\tilde{\mathbf{D}}$ is a diagonal matrix with $\tilde{\mathbf{D}}_{ii} = \sum_{j=1}^N \tilde{\mathbf{A}}_{ij}$.
In (4) each element of $\tilde{\mathbf{A}}$ is normalized by a factor equal to the square root of the
product of the degrees of the two corresponding nodes, namely $\mathbf{M}_{ij} = \frac{\tilde{\mathbf{A}}_{ij}}{\sqrt{\tilde{\mathbf{D}}_{ii} \tilde{\mathbf{D}}_{jj}}}$.

Unlike GCN, the DGCNN model (Zhang et al., 2018) calculates the aggre-

gation coefficients (\mathbf{M}_{ij}) as one-hop node transition probabilities ($\tilde{\mathbf{P}}_{ij}$) (see Section 3.2):

$$\mathbf{M} = \tilde{\mathbf{P}} = \tilde{\mathbf{D}}^{-1} \tilde{\mathbf{A}} \quad (5)$$

Note that in the DGCNN model, the aggregation coefficients are created by *normalizing the adjacency matrix using the degrees of the destination nodes*. Therefore, the neighboring nodes of a destination node have the same influence (i.e., weights) on the destination node, even though their popularity level (e.g., node degrees) may vary significantly. This is not desired as popular nodes are often connected with many other nodes, thus messages from the popular nodes are not as valuable as messages from nodes with lower level of popularity. The issue of popular nodes has been mentioned in (Do et al., 2017; Rahimi et al., 2015), where connections to these nodes are explicitly removed. In addition, we observe that in many text retrieval weighing schemes such as TF-IDF, the popular terms across documents are given smaller weights compared to less popular terms. This observation leads us to a new message passing formula, which we describe thoroughly in the next section.

4.1.2. The Proposed Graph Convolutional Layer

In this work, we propose a message passing scheme that makes use of the node transition probabilities, as in the DGCNN model (Zhang et al., 2018). Unlike the DGCNN, we *use the degrees of the source nodes* instead of the destination nodes for the normalization of the aggregation coefficients. Specifically, in our scheme, the aggregation coefficient matrix \mathbf{M} is calculated as:

$$\mathbf{M} = \tilde{\mathbf{A}}^T \tilde{\mathbf{D}}^{-1}. \quad (6)$$

In (6), \mathbf{M} contains one-hop node transition probabilities. \mathbf{M}_{ij} is the probability of transitioning from a source node j to a destination node i . As these probabilities are normalized using the degrees of the source nodes, we have $\sum_{i=1}^N \mathbf{M}_{ij} = 1, \forall j \in \{1, \dots, N\}$. It should be noted that in this case, $\sum_{j=1}^N \mathbf{M}_{ij} \neq 1$, which is different from the DGCNN model (as shown in (5)).

Substituting \mathbf{M} in (6) into the generic formulation in (3), we obtain the formulation for our graph convolutional layer as follows:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{A}}^T \tilde{\mathbf{D}}^{-1} \mathbf{H}^{(l)} \mathbf{W}). \quad (7)$$

We refer to this graph convolutional layer as the transition Probability based Graph **CONV**olutional layer, abbreviated as GPCONV (for ease of reference, we also use GCONV to refer to the graph convolutional layer proposed by (Kipf & Welling, 2016)). Intuitively, as the adjacency matrix is normalized by the degrees of source nodes, a node always has the same influence on its neighboring nodes. Specifically, a node’s message (feature vector) is disseminated to its neighbors with the same weight. As a result, a popular node (i.e., high degree) will have a smaller influence on its neighbors. As we analyzed earlier in Section 4.1.1, penalizing higher-degree nodes has been commonly enforced in various contexts, such as word embedding and node embedding. In the next section, we give a detailed example to illustrate this point.

4.1.3. Comparison with Existing Formulations

We denote by $\tilde{\mathbf{H}}^{(l+1)}$ the result of the aggregation step, i.e., the product of the aggregation coefficient matrix and the input representations, $\tilde{\mathbf{H}}^{(l+1)} = \mathbf{M}\mathbf{H}^{(l)}$. Thus:

$$\begin{aligned} \tilde{\mathbf{H}}_i^{(l+1)} &= \mathbf{M}_{i1} \mathbf{H}_1^{(l)} + \mathbf{M}_{i2} \mathbf{H}_2^{(l)} + \cdots + \mathbf{M}_{iN} \mathbf{H}_N^{(l)} \\ &= \sum_{j=1}^N \mathbf{M}_{ij} \mathbf{H}_j^{(l)}, \end{aligned} \quad (8)$$

$\forall i \in \{1, \dots, N\}$. Here, $\mathbf{H}_j^{(l)}$ and $\tilde{\mathbf{H}}_i^{(l+1)}$ are row vectors; $\mathbf{H}_j^{(l)}$ represents the representation (or feature) of node j at the l -th layer.

To illustrate the difference between the proposed graph convolutional layer’s formulation and the DGCNN model (Zhang et al., 2018), we consider an example of a graph containing four nodes, 1, 2, 3 and 4, as shown in Fig. 1. These nodes are associated with feature vectors $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ and $\mathbf{X}_4 \in \mathbb{R}^F$. We denote with $\tilde{\mathbf{P}} \in \mathbb{R}^{4 \times 4}$ the transition probability matrix, as calculated in (5),

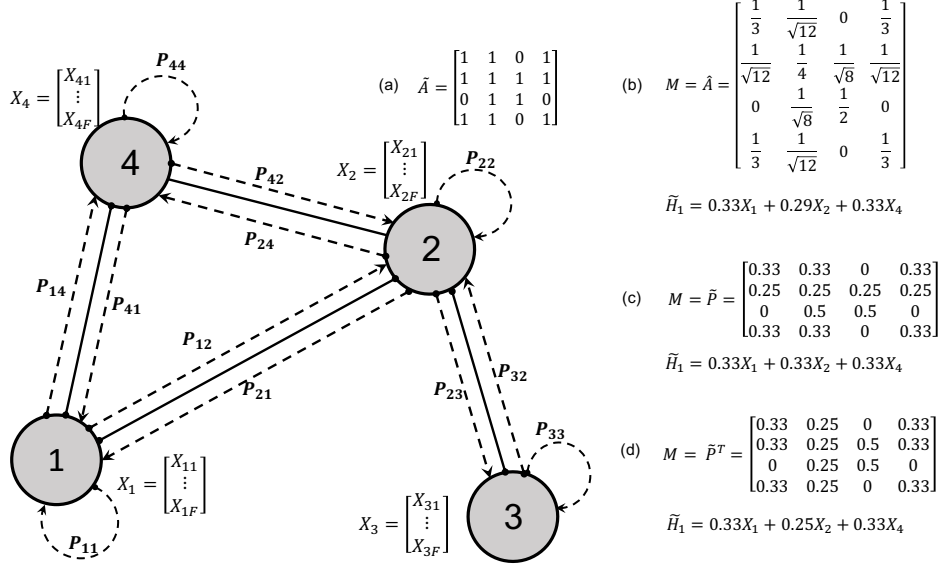


Figure 1: A graph with four nodes and four edges (solid lines) with directed transition probabilities indicated by dashed lines. (a) shows the adjacency matrix with self-connections added. (b), (c) and (d) show the equations to calculate the aggregated feature representation (i.e., \tilde{H}_1) for node 1, following the formulations in the GCN model (Kipf & Welling, 2016), the DGCNN model (Zhang et al., 2018) and the proposed scheme, respectively. We observe that in the proposed scheme, (i) the highest-degree node (i.e., node 2) is *less influential* in calculating \tilde{H}_1 than the other nodes, which are of lower degrees, and (ii) the overall influence of a node on their neighboring nodes are the same (i.e., entries in a column, where there exist corresponding connections, have the same value.).

and the *aggregate* step in the DGCNN model produces an aggregated feature representation for node 1 as:

$$\begin{aligned} \tilde{H}_1 &= \tilde{P}_{11}X_1 + \tilde{P}_{12}X_2 + \tilde{P}_{13}X_3 + \tilde{P}_{14}X_4 \\ &= 0.33X_1 + 0.33X_2 + 0.33X_4 \end{aligned} \quad (9)$$

Using the proposed scheme (i.e., (6)), we obtain:

$$\begin{aligned} \tilde{H}_1 &= \tilde{P}_{11}X_1 + \tilde{P}_{21}X_2 + \tilde{P}_{31}X_3 + \tilde{P}_{41}X_4 \\ &= 0.33X_1 + 0.25X_2 + 0.33X_4 \end{aligned} \quad (10)$$

Even though (10) has the same form as (9), the intuition behind the two is different. On the one hand, with our formulation, messages from each node

(i.e., source nodes 1, 2, 3, 4) to node 1 (destination node) are passed *following the probabilities of transitioning in the same directions*, namely, from the source nodes to destination node. On the other hand, in the DGCNN model, the same messages are passed *following the transitioning probabilities in the reverse directions*. We argue that the former is more intuitive and can lead to a more natural message passing scheme.

Another important distinction between the two formulations lie in the way the influence (i.e., weights) of the nodes are determined. As shown in Fig.1, the DGCNN model gives a weight of 0.33 to \mathbf{X}_2 in calculating $\tilde{\mathbf{H}}_1$ while the proposed scheme uses a weight of 0.25. Bear in mind that node 2 has the highest degree of 4 (self-loop added), we can conclude that the proposed scheme gives lower weight to nodes with higher degree. This leads to an effect similar to the TF-IDF weighting scheme as discussed in Section 4.1.1. In Section 5, we will empirically justify the benefits of the proposed formulation. Next, we will present how we build GCNN models, with the GPConv layer as the main building block, for node and graph classification tasks.

4.2. Node and Graph Classification Models

Using the proposed GPConv layer as a building block, we construct GCNN models for the *node* and *graph classification* tasks.

4.2.1. Node Classification Model

Our node classification model consists of L GPConv layers. The model's operation can be expressed by:

$$\mathbf{O} = \text{softmax}(\mathbf{M}\sigma(\dots\sigma(\mathbf{MXW}_1)\dots)\mathbf{W}_L). \quad (11)$$

In this (11), \mathbf{M} contains transition probabilities calculated using (6), $\mathbf{O} \in \mathbb{R}^{N \times C}$ are the predicted class probabilities where C is the number of classes. Note that the number of GPConv layers, L_{GPConv} , is a design choice. We refer to this model as the PGCN $_n$ model where P refers to transition probabilities and n stands for node classification.

4.2.2. Graph Classification Model

For graph the classification task, we need to predict a single label for the whole graph. To address this task, we design a model (similar to the node classification one) with a global pooling and several fully-connected layers added on top of the last GPCONV layer. The pooling operator is employed to produce a single representation for the whole graph, while the fully-connected layers and a softmax classifier are used to output the predicted label probabilities. There are several pooling techniques for graph classification, such as **SortPooling** (Zhang et al., 2018), **DiffPool** (Ying et al., 2018), **Top-K-Selection** (Gao & Ji, 2019), **max-pooling** (Zhang et al., 2018) and **mean-pooling** (Simonovsky & Komodakis, 2017; Monti et al., 2019; Bianchi et al., 2019). In our model, we use the global **mean-pooling** as it has been proven effective and is widely used for the task of graph classification (Monti et al., 2019; Bianchi et al., 2019). We refer to our graph classification model as the PGCN_g . Here, g stands for graph classification, which differentiates it from the proposed node classification model PGCN_n . The PGCN_g can be expressed by:

$$\mathbf{H} = \sigma(\mathbf{M}\sigma(\dots\sigma(\mathbf{M}\mathbf{X}\mathbf{W}_1)\dots)\mathbf{W}_L) \quad (12)$$

$$\mathbf{h} = \text{mean-pooling}(\mathbf{H}) \quad (13)$$

$$\mathbf{o} = \text{softmax}(\text{FC}(\mathbf{h})) \quad (14)$$

The (12), (13), (14) are written for one graph, so $\mathbf{H} \in \mathbb{R}^{N \times K}$ is a 2-D matrix, $\mathbf{h} \in \mathbb{R}^K$ and $\mathbf{o} \in \mathbb{R}^C$ are vectors. FC indicates the fully-connection part of the PGCN_g model. In practice, multiple graphs can be stacked together by concatenating the corresponding adjacency matrices diagonally and concatenating the feature matrices vertically. In that case, the pooling operation in (13) is applied graph-wise.

4.3. DropNode Regularizer

In this section, we present the proposed DropNode regularization method in detail. After that, we show how this method could be further incorporated into the node and graph classification models introduced in Section 4.2, and finally

we describe the connection between DropNode and the well-known dropout regularization method (Srivastava et al., 2014).

4.3.1. DropNode

300 The basic idea behind DropNode is to randomly sample sub-graphs from an input graph at each training iteration. This is achieved by dropping nodes following a Bernoulli distribution with a pre-defined probability $1 - p$, $p \in (0, 1)$. Such node dropping procedure can be seen as a *downsampling* operation, which reduces the dimension of the graph features by a factor of $1 - p$. To reconstruct
305 the original graph structure, each node dropping operation can be paired with an *upsampling* operation, which comes subsequently in the architecture of our GCNN models. The two operations are implemented as individual layers, which we refer to as the “downsampling” and “upsampling” layers.

A downsampling layer (which is the l -th layer in the model) takes an input
310 $\mathbf{H}^{(l)} \in \mathbb{R}^{N \times K_l}$, where N is the number of nodes and K_l the dimension of their representations. The layer randomly samples $N_l = \lfloor pN \rfloor$ rows in $\mathbf{H}^{(l)}$ to retain and remove the other $\lceil N(1 - p) \rceil$ rows ($\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ represent the floor and ceiling functions, respectively). Here, the value p is referred to as the *keep ratio*. The outputs of this layer are (i) a sub-matrix of $\mathbf{H}^{(l)}$, namely, $\mathbf{H}^{(l+1)} \in \mathbb{R}^{N_l \times K_l}$,
315 and (ii) a vector containing the indices of the rows in $\mathbf{H}^{(l)}$ that are retained. The second output will be used in a subsequent upsampling layer that is paired with this layer. Suppose that the aforementioned downsampling layer is paired with the upsampling layer, which is the k -th layer of the model where $k > l$, this upsampling layer takes as input a matrix $\mathbf{H}^{(k)} \in \mathbb{R}^{N_l \times K_k}$ and produces an
320 output matrix $\mathbf{H}^{(k+1)} \in \mathbb{R}^{N \times K_k}$. Each row in $\mathbf{H}^{(k)}$ is copied to a row in $\mathbf{H}^{(k+1)}$ according to the vector of indices obtained by the l -th layer. The rows in $\mathbf{H}^{(k+1)}$ that do not correspond to a row in $\mathbf{H}^{(l)}$ are filled in with zeros.

In our GCNN models, a downsampling layer follows a convolutional layer. Depending on the model design, upsampling layers may be used or not. Never-
325 theless, an upsampling layer must always correspond to a downsampling layer. Similar to dropout, the proposed DropNode method operates only during the

training phase. During the testing phase, all the nodes in the graph are used for prediction. We should note that in the case that the upsampling layers are not employed, the output of each downsampling layer needs to be scaled by a factor of $\frac{1}{p}$. This scaling operation is to maintain the same expected outputs for neurons in the subsequent layer during the training and testing phases (similar to dropout).

An important property of DropNode is that one or multiple sub-graphs are randomly sampled at each training iteration. Hence, the model does not see all the nodes during the training phase. As a result, the model should not rely on only a single prominent local pattern or on a small number of nodes, but to leverage information from all the nodes in the graph. The risk for the model to memorize the training samples, therefore, is reduced, avoiding over-fitting. In addition, the model is trained using multiple deformed versions of the original graphs. This can be considered as a data augmentation procedure, which is often used as an effective regularization method (DeVries & Taylor, 2017). On the other hand, the DropNode method reduces connectivity between nodes in the graph. Lower connectivity helps alleviate the smoothing of representation of the nodes when the GCNN model becomes deeper (Rong et al., 2019). As a result, features of nodes in different clusters will be more distinguishable, which could lead to improved performance on different downstream tasks for deep GCNN models.

4.3.2. Incorporating DropNode into the $PGCN_n$ Model

Figure 2 shows an example of how we incorporate DropNode into a $PGCN_n$ model for the node classification task. In this example, we employ three GPCONV layers and a pair of downsampling and upsampling layers. Given a graph G with input features \mathbf{X} , the model transforms \mathbf{X} through the first GPCONV layer into $\mathbf{H}^{(1)}$. Subsequently, the downsampling layer randomly drops a subset of rows of the transformed feature vectors $\mathbf{H}^{(1)}$. The remaining features $\mathbf{H}^{(2)}$ are fed into the second GPCONV layer. For the node classification task, it is essential to keep the same number of nodes at the output. To this end, an

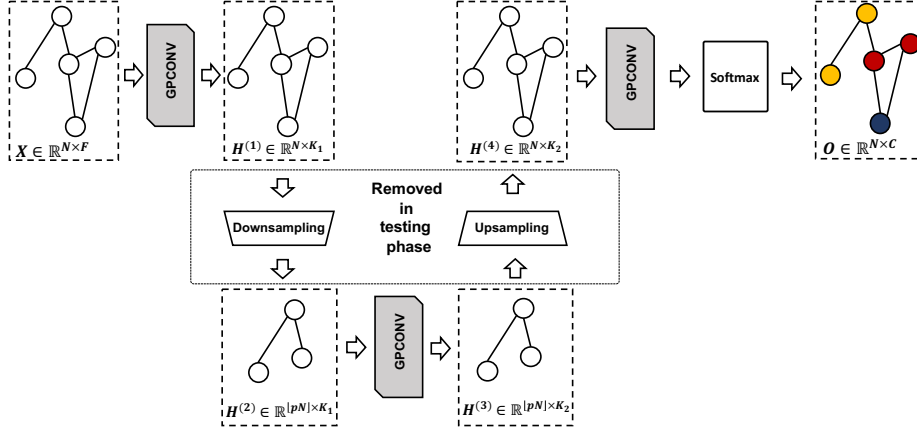


Figure 2: The structure of the $\text{PGCN}_n + \text{DropNode}$ model with two GPCONV layers and one pair of downsampling-upsampling layers for the node classification task.

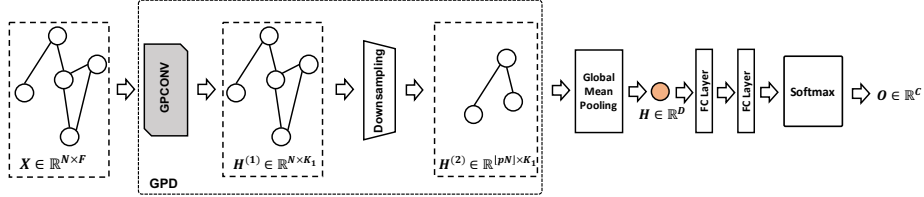


Figure 3: The $\text{PGCN}_g + \text{DropNode}$ model for graph classification. A GPCONV layer is combined with a downsampling layer to create a GPD block. Several GPD blocks can be stacked together to achieve better expressiveness power. The output of the final GPD block will be globally mean pooled making the final representation of the graph as denoted by \mathbf{H} . One or several fully-connected layer(s) (FC layers) and a softmax classifier are employed to predict the label of the graph.

upsampling layer is used to reconstruct the original graph structure. The output of this layer serves as input to the third GPCONV and a softmax classifier to produce class probabilities for all the nodes of the graph. It should be noted that the number of GPCONV layers L_{GPCONV} and downsampling-upsampling layer pairs L_{DU} are hyper-parameters of this model. We refer to this model as the $\text{PGCN}_n + \text{DropNode}$ model.

4.3.3. Incorporating DropNode into the PGCN_g Model

Figure 3 shows an example of how DropNode can be integrated with the PGCN_g model for the graph classification task. We refer to this model as PGCN_g + DropNode. The model has a block consisting of one GPCONV layer and one downsampling layer, which is abbreviated to “GPD” block. The GPD block is followed by a mean-pooling layer which produces a single representation vector for the whole graph. Then, two fully-connected layers act as a classifier on that representation vector. Similar to the PGCN_n + DropNode model (see Section 4.3.2), the numbers of GPD blocks L_{GPD} and fully-connected layers L_{FC} are hyper-parameters of the PGCN_g + DropNode model. In graph classification, the reconstruction of the graph structure is not needed (i.e., unlike the case of the node classification task). As such, it is not necessary to employ upsampling layers in the PGCN_g + DropNode model. As a result, the outputs of each downsampling layer need to be scaled by a factor of $\frac{1}{p}$ during training as mentioned earlier in Section 4.3.1.

4.3.4. Connection between DropNode and Dropout

DropNode and dropout both involve dropping activations in a layer of a neural network, yet, they are conceptually different. In DropNode, all the activations of a node are dropped. On the contrary, in dropout, the dropping is distributed across the nodes, so that on average, each node has a ratio of activations dropped. In addition, with DropNode, the structure of the graph is altered, while this does not happen in dropout. Implementation-wise, let $\mathbf{X} \in \mathbb{R}^{N \times K}$ be the input, dropout produces a matrix $\mathbf{X}' \in \mathbb{R}^{N \times K}$ of the same dimensions as \mathbf{X} in which some entries are randomly set to zero; whereas, DropNode produces a sub-matrix $\mathbf{X}'' \in \mathbb{R}^{\lfloor pN \rfloor \times K}$ where $p \in (0, 1)$. Figures 4, 5 show the difference between dropout and DropNode.

$$\begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & X_{13} & X_{14} \\ \hline X_{21} & X_{22} & X_{23} & X_{24} \\ \hline X_{31} & X_{32} & X_{33} & X_{34} \\ \hline X_{41} & X_{42} & X_{43} & X_{44} \\ \hline X_{51} & X_{52} & X_{53} & X_{54} \\ \hline \end{array} \times \begin{array}{|c|c|c|c|} \hline 1 & 1 & & 1 \\ \hline 1 & & 1 & 1 \\ \hline 1 & 1 & & 1 \\ \hline 1 & 1 & 1 & \\ \hline & 1 & 1 & 1 \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & 0 & X_{14} \\ \hline X_{21} & 0 & X_{23} & X_{24} \\ \hline X_{31} & X_{32} & 0 & X_{34} \\ \hline X_{41} & X_{42} & X_{43} & 0 \\ \hline 0 & X_{52} & X_{53} & X_{54} \\ \hline \end{array}$$

$\mathbf{X} \in \mathbb{R}^{5 \times 4}$
 $p = 0.75$
 $\mathbf{X}' \in \mathbb{R}^{5 \times 4}$

Figure 4: Dropout applied to feature matrix \mathbf{X} . Several neural units corresponding to elements on each row of \mathbf{X} are randomly deactivated during the training phase. This operation produces matrix \mathbf{X}' of the same shape as the input.

$$\begin{array}{|c|c|c|c|} \hline X_{11} & X_{12} & X_{13} & X_{14} \\ \hline X_{21} & X_{22} & X_{23} & X_{24} \\ \hline X_{31} & X_{32} & X_{33} & X_{34} \\ \hline X_{41} & X_{42} & X_{43} & X_{44} \\ \hline X_{51} & X_{52} & X_{53} & X_{54} \\ \hline \end{array} \times \begin{array}{|c|} \hline \\ \hline 1 \\ \hline 1 \\ \hline \\ \hline 1 \\ \hline \end{array} = \begin{array}{|c|c|c|c|} \hline X_{21} & X_{22} & X_{23} & X_{24} \\ \hline X_{31} & X_{32} & X_{33} & X_{34} \\ \hline X_{51} & X_{52} & X_{53} & X_{54} \\ \hline \end{array}$$

$\mathbf{X} \in \mathbb{R}^{5 \times 4}$
 $p = 0.75$
 $\mathbf{X}'' \in \mathbb{R}^{3 \times 4}$

Figure 5: DropNode applied to feature matrix \mathbf{X} . Unlike dropout, some rows are randomly eliminated, leading to matrix \mathbf{X}'' with smaller number of rows.

5. Experiments

5.1. Datasets

In our experiments, we consider both node and graph classification tasks. For node classification, we use three benchmark citation network datasets¹, namely, CORA, CITESEER and PUBMED (Sen et al., 2008). Graphs are created for these datasets by considering scientific papers as nodes and references between the papers as edges. Each node is represented by a bag-of-words feature vector extracted from the corresponding document. In these datasets, all the nodes are labeled. Following Kipf et al. (Kipf & Welling, 2016) and Veličković et al. (Veličković et al., 2017), we consider the graphs of CORA, CITESEER and PUBMED *undirected*, although the references are actually directed. The description of the considered datasets for node classification is presented in

¹<https://lincs.soe.ucsc.edu/data>

Table 1: Datasets for the node classification tasks. $+/-$ indicates whether the corresponding features are available or unavailable. The numbers in parentheses denote the dimensionality of the corresponding feature vectors.

	CORA	CITESEER	PUBMED
# Nodes	2708	3327	19,717
# Edges	5429	4732	44,338
# Labels	7	6	3
Node Attr.	+(1433)	+(3703)	+(500)
Edge Attr.	—	—	—

Table 2: Datasets for graph classification task. $+/-$ indicates whether the corresponding features are available or unavailable. The numbers in parentheses denote the dimensionality of the corresponding feature vectors.

	PROTEINS	D&D	ENZYMES	MUTAG	NCI1
# Graph	1113	1178	600	188	4110
# Label	2	2	6	2	2
# Avg. Node	39.06	284.32	32.63	17.93	29.87
# Avg. Edge	72.82	715.66	62.14	19.79	32.30
Node Label	+	+	+	+	+
Edge Label	—	—	—	+	—
Node Attr.	+(29)	—	+(18)	—	—
Edge Attr.	—	—	—	—	—

Table 1.

Concerning the graph classification task, we employ the following datasets²: the bioinformatics datasets, namely ENZYMES, PROTEINS, D&D, MUTAG; the scientific collaboration dataset COLLAB (Kersting et al., 2016); the chemical
405 compound dataset NCI1. In the bioinformatics datasets, each graph represents a biological structure. For the COLLAB dataset, a graph represents an ego-network of researchers who have collaborated with each other (Gomez et al., 2017). The NCI1 dataset represents the activity against non-small cell lung

²<https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets>

cancer (Gomez et al., 2017). The description of these datasets is presented in
410 Table 2.

5.2. Experimental Setup

For both node and graph classification tasks, we employ classification accuracy as the performance metric. Concerning the node classification task, similar to (Kipf & Welling, 2016; Veličković et al., 2017), we consider the *transductive* experimental setting (see Section 3). We employ the standard train/validation/test
415 set splits of all the considered datasets for node classification to guarantee a fair comparison with prior works (Kipf & Welling, 2016; Veličković et al., 2017; Gao & Ji, 2019). That is, (140/500/1000) nodes are used for the CORA dataset, (120/500/1000) nodes are used for the CITESEER dataset and (60/500/1000)
420 nodes are used for the PUBMED dataset. It is worth mentioning that the amount of labelled nodes is much smaller than the amount of test nodes, which makes the task highly challenging. Similar to (Veličković et al., 2017; Gao & Ji, 2019; Veličković et al., 2018), we report the mean and standard deviation of the results over 100 runs with random weight initialization.

Concerning the graph classification task, following existing works (Ying et al.,
425 2018; Zhang et al., 2018), we employ a 10-fold cross validation procedure and report the average accuracy over the folds. Among the graph classification datasets, only PROTEINS and ENZYMES provide node features (see Table 2), which can be used directly as input to the proposed models. For the rest of the
430 datasets, we use the degree and the labels of the nodes as features.

We compare the performance of our models with state-of-the-art baseline models. Specifically, for the node classification task, the selected baselines are the GCN (Kipf & Welling, 2016), GAT (Veličković et al., 2017), GraphSAGE (Hamilton et al., 2017), DGI (Veličković et al., 2018), GMNN (Qu et al., 2019), Graph
435 U-Net (Gao & Ji, 2019), DGCNN (Zhang et al., 2018), and DropEdge (Rong et al., 2019) models. The DGCNN model is originally designed for graph classification. In order to use this model for node classification, we employ only its message passing mechanism (see Section 4.1.1), ignoring its pooling part. For the

graph classification task, the Graph U-Net (Gao & Ji, 2019), DGCNN (Zhang et al., 2018), DiffPool (Ying et al., 2018), GraphSAGE (Hamilton et al., 2017), CapsGNN (Xinyi & Chen, 2019) and SAGPool (Lee et al., 2019) models are selected. In addition, following (Luzhnica et al., 2019), we employ two simple baselines including a fully-connected neural network with two hidden layers denoted by FCN, and a combination of FCN with one graph convolutional layer (GCONV) from the popular GCN model (Kipf & Welling, 2016) (see Section 4.1.1) denoted as GCN + 2FC. For each baseline model and a benchmark dataset, we select the best results reported in the corresponding paper (if available). Otherwise, we collect the results using either the implementations released by the corresponding authors or self-implemented source code.

5.3. Hyperparameter Settings

Hyperparameters of the proposed models are found empirically via tuning. For the node classification model PGCN_n , we use two GPCONV layers ($L_{\text{GPCONV}} = 2$). This choice also follows the best configuration suggested in (Kipf & Welling, 2016). Each GPCONV layer has a hidden dimension of 64. In addition, dropout is added after each GPCONV layer with a dropping rate of 0.7. When DropNode is used (i.e., the $\text{PGCN}_n + \text{DropNode}$ model), we employ three GPCONV layers ($L_{\text{GPCONV}} = 3$), also with a hidden dimension of 64, and a pair of downsampling-upsampling layers. Compared to the PGCN_n model, one additional GPCONV layer is added between the downsampling and upsampling layers. The downsampling layer of DropNode has the keep ratio p selected in such a way that 200 nodes are retained for all the datasets. We do not employ dropout for $\text{PGCN}_n + \text{DropNode}$ as the DropNode has already had the regularization effect on the considered model. We train the two models with learning rates of 0.01 and 0.001, respectively.

For our graph classification models, namely, PGCN_g and $\text{PGCN}_g + \text{DropNode}$, we employ one GPCONV layer ($L_{\text{GPCONV}} = 1$) and a single GPD block ($L_{\text{GPD}} = 1$), respectively. Both models use two fully-connected layers ($L_{\text{FC}=2}$). The GPCONV and the fully-connected layers have 512 hidden units each. We

Table 3: Node classification results in terms of the accuracy evaluation metric (%). We report the mean and standard deviation of the accuracy over 100 runs. The bold font indicates the best performance. Our models include PGCN_n and $\text{PGCN}_n + \text{DropNode}$. In addition, we apply DropNode to the common GCN model (Kipf & Welling, 2016), referred to as $\text{GCN}_n + \text{DropNode}$. The asterisk (*) indicates that the result is obtained by using our own implementation.

Method	CORA	CITeseer	PUBMED
GCN + DropEdge (2019)	82.80	72.30	79.60
GMNN (2019)	83.7	72.9	81.8
GCN (2016)	81.9 ± 0.7	70.5 ± 0.8	78.9 ± 0.5
DGCNN* (2018)	81.4 ± 0.5	69.8 ± 0.7	78.1 ± 0.4
GAT (2017)	83.0 ± 0.7	72.5 ± 0.7	79.0 ± 0.3
Graph U-Net (2019)	84.4 ± 0.6	73.2 ± 0.5	79.6 ± 0.2
DGI (2018)	82.3 ± 0.6	71.8 ± 0.7	76.8 ± 0.6
PGCN_n	81.7 ± 0.5	70.6 ± 0.7	78.4 ± 0.4
$\text{GCN}_n + \text{DropNode}$	84.6 ± 1.0	74.3 ± 0.5	82.7 ± 0.2
$\text{PGCN}_n + \text{DropNode}$	85.1 ± 0.7	74.3 ± 0.6	83.0 ± 0.3

employ dropout after each layer with a dropping ratio of 0.5 in the PGCN_g model. Similar to the $\text{PGCN}_n + \text{DropNode}$ model, we do not use dropout for $\text{PGCN}_g + \text{DropNode}$. For the $\text{PGCN}_g + \text{DropNode}$, the keep ratio p is set to $p = 0.75$, which is much higher than that used for the node classification models. This is due to the fact that in the considered graph classification datasets, the graph sizes are much smaller than those in the node classification datasets (see Table 2). We train both models using a small learning rate of 0.0001.

5.4. Experimental Results

5.4.1. Node Classification

The node classification results of different models are reported in Table 3. The reported results include the mean and standard deviation of classification accuracy over 100 runs. The results show that the PGCN_n and GCN (Kipf & Welling, 2016) models achieve higher accuracy compared to the DGCNN model. This can be attributed to better message passing schemes giving smaller

weights to popular nodes presented in Section 4.1.2 as these three models
 have similar configuration including number of graph convolutional layers and
 485 number of hidden units in each layer. Furthermore, the PGCN_n model achieves
 marginally better performance compared to the popular GCN model on the
 CITESEER dataset, reaching 70.6% compared to 70.5% obtained by the GCN
 model. Nevertheless, the PGCN_n model performs slightly worse than the GCN
 model on the CORA and PUBMED datasets, amounting to a 0.2% and 0.5% drop
 490 in terms of accuracy. This can be explained by the fact that the variance of the
 node degree distribution of the CORA ($\sigma_{\text{degree}} = 5.23$) and PUBMED datasets
 ($\sigma_{\text{degree}} = 7.43$) are higher than that of the CITESEER dataset ($\sigma_{\text{degree}} = 3.38$).
 Recall that the PGCN_n model assigns much lower weights on higher degree nodes.
 Therefore, the PGCN_n model might perform slightly worse compared to GCN
 495 on datasets with high degree imbalance. On the other hand, on datasets with
 balanced node degrees, such as CITESEER, the PGCN_n model performs better
 than GCN. In addition, compared to the other baseline models, the PGCN_n
 model achieves lower accuracy on CORA, whereas it produces comparable results
 on CITESEER and PUBMED. When DropNode is used, it consistently improves
 500 the performance of all considered models. Specifically, $\text{PGCN}_n + \text{DropNode}$ and
 $\text{GCN}_n + \text{DropNode}$ significantly improve the performance of PGCN_n and GCN_n
 by around 4 percentage points of accuracy. In particular, $\text{GCN}_n + \text{DropNode}$
 can reach 84.6%, 74.3% and 82.7% while $\text{PGCN}_n + \text{DropNode}$ achieves the
 best performance with 85.1%, 74.3% and 83.0% on the CORA, CITESEER
 505 and PUBMED datasets, respectively. It is worth recalling that in our setting,
 the number of training examples is much smaller compared to the number of
 testing examples. By using DropNode, deformed versions of the underlying
 graph are created during each training epoch. In other words, DropNode acts
 as an augmentation technique on the training data which leads to an increased
 510 performance.

Table 4: Graph classification result in terms of percent (%). FCN stands for fully-connected neural network (2 FC layers). N/A stands for not available. Daggers mean the results are produced by running the code of the authors on corresponding datasets (the results are not available in the original paper).

Method	PROTEINS	DD	ENZYMES	MUTAG	NCI1
Diff-Pool (GraphSAGE) (2018)	70.48	75.42	54.25	N/A	N/A
Diff-Pool (Soft Assign) (2018)	76.25	80.64	62.53	88.89 [†]	80.36 [†]
Graph U-Net (2019)	77.68	82.43	48.33 [†]	86.76 [†]	72.12 [†]
CapsGNN (2019)	76.28	75.38	54.67	86.67	78.35
DGCNN (2018)	75.54	79.37	46.33 [†]	85.83	74.44
SAGPool _g (2019)	70.04	76.19	N/A	N/A	74.18
SAGPool _h (2019)	71.86	76.45	N/A	N/A	67.45
FCN (2FC)	74.68	75.47	66.17	87.78	69.69
GCN + 2FC	74.86	75.64	66.45	86.11	75.90
PGCN _g + 2FC	75.13	78.46	66.17	85.55	75.84
GCN _g + DropNode	76.58	79.32	69.00	87.27	79.03
PGCN _g + DropNode	77.21	80.69	70.50	89.44	81.11

5.4.2. Graph Classification

The results for graph classification are given in Table 4. In addition to the models mentioned in Section 4.2 and Section 4.3.1, we also provide the result produced by a simple fully-connected neural network, denoted by FCN, with two hidden layers; each has size of 512 units. We observe that the simple FCN can achieve high classification accuracy compared to presented strong baselines on some datasets. For instance, FCN obtains 74.68% on PROTEINS, which is around 4 percentage points higher than the performance of GraphSAGE and SAGPool_g, and approximately 3 percentage points higher than SAGPool_h. The good performance of structure-blind fully-connect neural network has been reported by (Luzhnica et al., 2019), which is also confirmed in our work. By adding a GCONV or GPCONV layer on top of the FCN model (GCN + 2FC, PGCN_g + 2FC) the accuracy on PROTEINS, DD and ENZYMES is marginally improved while the accuracy on NCI1 is improved by 6 percentage points. This is because the GCONV / GPCONV layers are able to exploit the graph structure of

the considered bioinformatics datasets. By using DropNode, the performance of our models is further improved. Specifically, our model with a single GPCONV layer, two FC layers and DropNode (i.e., $\text{PGCN}_g + \text{DropNode}$) outperforms all the baselines on PROTEINS, ENZYMES, MUTAG and NCI1, except for DD where our models perform slightly worse compared to the Graph U-Net model. Even we do not outperform the Graph U-Net on the DD dataset, it is clear that DropNode improves the PGCN_g by more than 2% accuracy point. This again confirms the consistency of DropNode in improving GCNN models.

5.5. The Effect of DropNode on Deep GCNNs

In this section, we investigate the effect of DropNode on deeper graph convolutional models. Specifically, we run our best model, which is comprised of many GPCONV layers with and without DropNode for node classification on the CORA and CITESEER datasets. The number of GPCONV layers L_{GPCONV} is set to 5, 7 and 9; each GPCONV layer has a hidden dimension of 64. The numbers of nodes that are kept in each case are shown in Table 5. The rest of the parameters have the same values as presented in Section 5.3. The corresponding results are presented in Table 6.

We observe that the performance of the GCN and PGCN_n models decreases significantly when the number of hidden layers increases. Specifically, 9GPCONV-layer GCN produces an accuracy score of only 13% on CORA and 22.2% on CITESEER while similar performance is produced by a PGCN_n model with 9GPCONV layers. This can be explained by the fact that (i) deep GCN / PGCN_n models have many more parameters compared to the shallow ones, which are prone to over-fitting, and (ii) the deep models suffer from over-smoothing (Li et al., 2018; Chen et al., 2019), which results in indistinguishable node representations for the different classes. By applying DropNode on both models, the classification accuracy is improved significantly, especially in the case that 7 and 9 layers are used. This is because the effects of over-fitting and over-smoothing are alleviated.

Table 5: Number of nodes kept for downsampling layers. 3GPConv indicates that there are three GPConv layers used. #DL stands for downsampling layer dimensionality. “—” means not applicable.

	3GPConv	5GPConv	7GPConv	9GPConv
#DL 1	200	200	200	200
#DL 2	—	150	150	150
#DL 3	—	—	100	100
#DL 4	—	—	—	50

Table 6: Accuracy (%) of deep GCNNs with DropNode integrated.

	3 layers		5 layers		7 layers		9 layers	
	CORA	CITESEER	CORA	CITESEER	CORA	CITESEER	CORA	CITESEER
GCN	79.1	69.0	78.8	61.8	46.2	23.0	13.0	22.2
PGCN	79.8	69.0	77.6	64.4	52.7	35.4	13.0	25.10
GCN _n + DropNode	84.60	74.30	80.80	72.10	71.80	70.40	51.20	66.70
PGCN _n + DropNode	85.10	74.30	81.40	72.30	75.10	70.70	49.60	66.90

6. Conclusion

In this work, we have proposed a new graph message passing mechanism, which leverages the transition probabilities of nodes in a graph, for graph convolutional neural networks (GCNNs). The proposed message passing mechanism is simple, however, it achieves good performance for the common tasks of node and graph classification. Additionally, we have introduced a novel technique termed DropNode for regularizing the GCNNs. The DropNode regularization technique can be integrated into existing GCNN models leading to noticeable improvements on the considered tasks. Furthermore, it has been shown that DropNode works well under the condition that the number of labelled examples is limited, which is useful in many real-life applications when it is normally hard and expensive to collect a substantial amount of labelled data. Our future work

will focus on generalizing the proposed method on large graphs, e.g., reducing the computational cost of re-computing intermediate adjacency matrices. In addition, as our method is general, it could be applied to a wide range of applications involving graph-structured data such as social media or Internet-of-Things data.

References

- Atwood, J., & Towsley, D. (2016). Diffusion-convolutional neural networks. In *Advances in neural information processing systems* (pp. 1993–2001).
- Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J. et al. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems* (pp. 4502–4510).
- Bianchi, F. M., Grattarola, D., Livi, L., & Alippi, C. (2019). Graph neural networks with convolutional arma filters. *arXiv preprint arXiv:1901.01343*, .
- Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, .
- Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., & Sun, X. (2019). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *arXiv preprint arXiv:1909.03211*, .
- Chen, J., Ma, T., & Xiao, C. (2018). Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv:1801.10247*, .
- Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems* (pp. 3844–3852).
- DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, .
- Do, T. H., Nguyen, D. M., Tsiligianis, E., Aguirre, A. L., La Manna, V. P., Pasveer, F., Philips, W., & Deligiannis, N. (2019). Matrix completion with

- variational graph autoencoders: Application in hyperlocal air quality inference.
In *IEEE International Conference on Acoustics, Speech and Signal Processing*
595 *(ICASSP)* (pp. 7535–7539).
- Do, T. H., Nguyen, D. M., Tsiligianni, E., Cornelis, B., & Deligiannis, N. (2017).
Multiview deep learning for predicting twitter users’ location. *arXiv preprint*
arXiv:1712.08091, .
- Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T.,
600 Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs
for learning molecular fingerprints. In C. Cortes, N. D. Lawrence, D. D. Lee,
M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing*
Systems 28 (pp. 2224–2232). Curran Associates, Inc.
- Gao, H., & Ji, S. (2019). Graph u-nets. *arXiv preprint arXiv:1905.05178*, .
- 605 Gao, H., Wang, Z., & Ji, S. (2018). Large-scale learnable graph convolutional
networks. In *Proceedings of the 24th ACM SIGKDD International Conference*
on Knowledge Discovery & Data Mining (pp. 1416–1424).
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017).
Neural message passing for quantum chemistry. In *Proceedings of the 34th*
610 *International Conference on Machine Learning-Volume 70* (pp. 1263–1272).
JMLR. org.
- Gomez, L. G., Chiem, B., & Delvenne, J.-C. (2017). Dynamics based features
for graph classification. *arXiv preprint arXiv:1705.10817*, .
- Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning
615 on large graphs. In *Advances in Neural Information Processing Systems* (pp.
1024–1034).
- Henaff, M., Bruna, J., & LeCun, Y. (2015). Deep convolutional networks on
graph-structured data. *arXiv preprint arXiv:1506.05163*, .

- Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular
620 graph convolutions: moving beyond fingerprints. *Journal of computer-aided
molecular design*, 30(8), 595–608.
- Kersting, K., Kriege, N. M., Morris, C., Mutzel, P., & Neumann, M.
(2016). Benchmark data sets for graph kernels. <http://graphkernels.cs.tu-dortmund.de>.
- 625 Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph
convolutional networks. *arXiv preprint arXiv:1609.02907*, .
- Lee, J., Lee, I., & Kang, J. (2019). Self-attention graph pooling. *arXiv preprint
arXiv:1904.08082*, .
- Leskovec, J., & Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings
630 of the 12th ACM SIGKDD international conference on Knowledge discovery
and data mining* (pp. 631–636). ACM.
- Li, Q., Han, Z., & Wu, X.-M. (2018). Deeper insights into graph convolutional
networks for semi-supervised learning. In *Thirty-Second AAAI Conference on
Artificial Intelligence*.
- 635 Li, Y., Tarlow, D., Brockschmidt, M., & Zemel, R. (2015). Gated graph sequence
neural networks. *arXiv preprint arXiv:1511.05493*, .
- Luzhnica, E., Day, B., & Liò, P. (2019). On graph classification networks,
datasets and baselines. *arXiv preprint arXiv:1905.04682*, .
- Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019).
640 Fake news detection on social media using geometric deep learning. *arXiv
preprint arXiv:1902.06673*, .
- Niepert, M., Ahmed, M., & Kutzkov, K. (2016). Learning convolutional neural
networks for graphs. In *International conference on machine learning* (pp.
2014–2023).

- 645 Ortega, A., Frossard, P., Kovačević, J., Moura, J. M., & Vandergheynst, P. (2018). Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106(5), 808–828.
- Qi, X., Liao, R., Jia, J., Fidler, S., & Urtasun, R. (2017). 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5199–5208).
650
- Qu, M., Bengio, Y., & Tang, J. (2019). Gmn: Graph markov neural networks. *arXiv preprint arXiv:1905.06214*, .
- Quek, A., Wang, Z., Zhang, J., & Feng, D. (2011). Structural image classification with graph neural networks. In *2011 International Conference on Digital Image Computing: Techniques and Applications* (pp. 416–421). IEEE.
655
- Rahimi, A., Cohn, T., & Baldwin, T. (2015). Twitter user geolocation using a unified text and network prediction model. *arXiv preprint arXiv:1506.08259*, .
- Rong, Y., Huang, W., Xu, T., & Huang, J. (2019). Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations*.
660
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234–241). Springer.
- 665 Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1), 1–8.
- Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI magazine*, 29(3), 93–93.
- 670 Simonovsky, M., & Komodakis, N. (2017). Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3693–3702).

- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.
- Teney, D., Liu, L., & van Den Hengel, A. (2017). Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*, .
- Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2018). Deep graph infomax. *arXiv preprint arXiv:1809.10341*, .
- Xinyi, Z., & Chen, L. (2019). Capsule graph neural network. In *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=Byl8BnRcYm>.
- Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7370–7377). volume 33.
- Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems* (pp. 4800–4810).
- Zhang, M., Cui, Z., Neumann, M., & Chen, Y. (2018). An end-to-end deep learning architecture for graph classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., & Sun, M. (2018). Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, .