# Exploiting Edge Features for Graph Neural Networks

Liyu Gong[1] Qiang Cheng *[1,2]
[1] Institute for Biomedical Informatics, University of Kentucky, Lexington, USA
[2] Department of Computer Science, University of Kentucky, Lexington, USA
{liyu.gong, Qiang.Cheng}@uky.edu

## Abstract

*Edge features contain important information about graphs. However, current state-of-the-art neural network models designed for graph learning,* e.g.*, graph convolutional networks (GCN) and graph attention networks (GAT), inadequately utilize edge features, especially multi-dimensional edge features. In this paper, we build a new framework for a family of new graph neural network models that can more sufficiently exploit edge features, including those of undirected or multi-dimensional edges. The proposed framework can consolidate current graph neural network models,* e.g.*, GCN and GAT. The proposed framework and new models have the following novelties: First, we propose to use doubly stochastic normalization of graph edge features instead of the commonly used row or symmetric normalization approaches used in current graph neural networks. Second, we construct new formulas for the operations in each individual layer so that they can handle multi-dimensional edge features. Third, for the proposed new framework, edge features are adaptive across network layers. Fourth, we propose to encode edge directions using multi-dimensional edge features. As a result, our proposed new framework and new models are able to exploit a rich source of graph edge information. We apply our new models to graph node classification on several citation networks, whole graph classification, and regression on several molecular datasets. Compared with the current state-of-the-art methods,* i.e.*, GCNs and GAT, our models obtain better performance, which testify to the importance of exploiting edge features in graph neural networks.*

## 1. Introduction

Deep neural networks have become one of the most successful machine learning techniques in recent years. In many important problems, they achieve state-of-the-art performance, *e.g.*, convolutional neural networks (CNN) [19] in image recognition, and recurrent neural networks (RNN) [12] and Long Short Term Memory (LSTM) [14] in natural language processing. In real world, many problems can be naturally modeled with graphs rather than conventional tables, grid type images, or time sequences. Generally, a

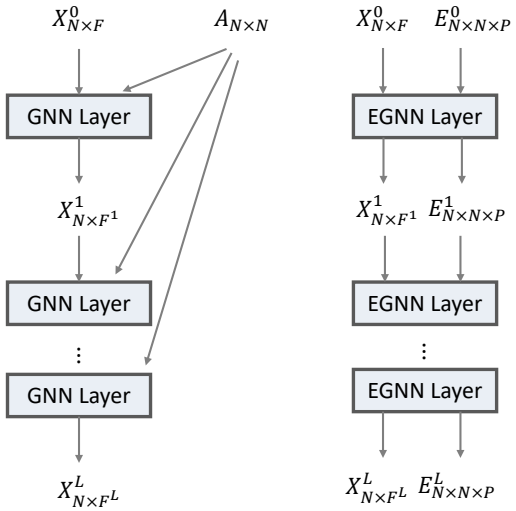



Figure 1: Schematic illustration of the proposed edge enhanced graph neural network (EGNN) architecture (right), compared with the original graph neural network (GNN) architecture (left). A GNN layer could be a GCN layer, or a GAT layer, while an EGNN layer is an edge enhanced counterpart of it. EGNN differs from GNN structurally in two folds. Firstly, the adjacency matrix $A$ in GNN is either a binary matrix that indicates merely the neighborhood of each node and is used in GAT layers, or a nonnegative-valued matrix that has one dimensional edge features and is used in GCN layers; in contrast, EGNN uses the multi-dimensional nonnegative-valued edge features represented as a tensor $E$ which may exploit multiple attributes associated with each edge. Secondly, in GNN the same original adjacency matrix $A$ is fed to every layer; in contrast, the edge features in EGNN are adapted at each layer before being fed to next layer.

*Corresponding author.

graph contains nodes and edges, where nodes represent entities in real world, and edges represent interactions or relationships between entities. For example, a social network naturally models users as nodes and friendship relationships as edges. For each node, there is often an associated feature vector describing it, *e.g*., a user's profile in a social network. Similarly, each edge is also often associated with features depicting relationship strengths or other properties. Due to their complex structures, a challenge in learning on graphs is to find effective ways to incorporate different sources of information contained in graphs into computational models such as neural networks.

Recently, several neural network models have been developed for graph learning, which obtain better performance than traditional techniques. Inspired by graph Fourier transform, Defferrard *et al*. [11] propose a graph convolution operation as an analogue to standard convolutions used in CNN. Just like the convolution operation in image spatial domain is equivalent to multiplication in the frequency domain, convolution operators defined by polynomials of a graph Laplacian is equivalent to filtering in the graph spectral domain. Particularly, by applying Chebyshev polynomials to the graph Laplacian, spatially localized filtering is obtained. Kipf *et al*. [18] approximate the polynomials using a re-normalized first-order adjacency matrix to obtain comparable results on graph node classification tasks. Those graph convolutional networks (GCNs) [11][18] combine graph node features and graph topological structural information to make predictions. Velickovic *et al*. [27] adopt attention mechanism into graph learning, and propose a graph attention network (GAT). Unlike GCNs, which use a fixed or learnable polynomial of Laplacian or adjacency matrix to aggregate (filter) node information, GAT aggregates node information by using an attention mechanism on graph neighborhoods. The essential difference between GAT and GCNs is stark: In GCNs the weights for aggregating (filtering) neighbor nodes are defined by the graph topological structure, which is independent of node contents; in contrast, weights in GAT are a function of node contents due to the attention mechanism. Empirical results on graph node classification show that the adaptiveness of GAT makes it more effective to fuse information from node features and graph topological structures.

One major problem in the current GNN models, such as GAT and GCNs, is that edge features are not fully incorporated. In GAT, graph topological information is injected into the model by forcing the attention coefficient between two nodes to zero if they are not connected. Therefore, the edge information used in GAT is only the indication about whether there is an edge or not, *i.e*., connectivities. However, graph edges are often in possession of rich information like strengths, types, etc. Instead of being a binary indicator variable, edge features could be continuous, *e.g*., strengths, or multi-dimensional. GCNs can utilize one-dimensional real-valued edge features, *e.g*., edge weights, but the edge features are restricted to be one-dimensional. Properly addressing this problem is likely to benefit many graph learning problems. Another problem of GAT and GCNs is that each GAT or GCN layer filters node features based on the original adjacency matrix that is given as an input. The original adjacency matrix is likely to be noisy and not optimal, which will limit the effectiveness of the filtering operation.

In this paper, we address the above problems by proposing new GNN models to more adequately exploit edge information, which naturally enhance current GCNs and GAT models. Our models construct different formulas from those of GCNs and GAT, so that they are capable of exploiting multi-dimensional edge features. Also our new models can exploit one-dimensional edge features more effectively by making them adaptive across network layers. Moreover, our models leverage doubly stochastic normalization to augment the GCNs and GAT models that use ordinary row or symmetric edge normalization. Doubly stochastic matrices have nice properties that can facilitate the use of edges.

We conduct experiments on several citation network datasets and molecular datasets. For citation networks, we encode directed edges as three dimensional edge feature vectors. For molecular datasets, different atom bond types are naturally encoded as multi-dimensional edge attributes. By leveraging those multi-dimensional edge features our methods outperform current state-of-the-art approaches. The results confirm that edge features are important for graph learning, and our proposed EGNN models are effective incorporating edge features.

As a summary, the novelties of our proposed EGNN model include the following:

- A new framework for adequately exploiting multi-dimensional edge features. Our new framework is able to incorporate multi-dimensional positive-valued edge features. It eliminates the limitation of GAT which can handle only binary edge indicators and the limitation of GCNs which can handle only one dimensional edge features.

- Doubly stochastic edge normalization. We propose to normalize edge feature matrices into doubly stochastic matrices which show improved performance in denoising [29].

- Attention based edge adaptiveness across neural network layers. We design a new graph network architecture which can not only filter node features but also adapt edge features across layers. Leveraging this new architecture, in our models the edge features are adaptive to both local contents and the global layers when passing through the layers of the neural network.

- Multi-dimensional edge features for directed edges. We propose a method to encode edge directions as multi-dimensional edge features. Therefore, our EGNN can effectively learn on directed graphs.

The rest of this paper is organized as follows: Section 2 briefly reviews the related works. Details of the proposed EGNN architecture and two types of proposed EGNN layers are described in Section 3. Section 4 presents the experimental results, and Section 5 concludes the paper.

## 2. Related works

A critical challenge in graph learning is the complex non-Euclidean structure of graph data. To address this challenge, traditional machine learning approaches extract graph statistics (*e.g.*, degrees) [5], kernel functions [28][24] or other hand-crafted features which measure local neighborhood structures. Those methods lack flexibility in that designing sensible hand-crafted features is time consuming and extensive experiments are needed to generalize to different tasks or settings. Instead of extracting structural information or using hand-engineered statistics as features of the graph, graph representation learning attempts to embed graphs or graph nodes in a low-dimensional vector space using a data-driven approach. One kind of embedding approaches are based on matrix-factorization, *e.g.*, Laplacian Eigenmap (LE) [4], Graph Factorization (GF) algorithm [2], GraRep [7], and HOPE [21]. Another class of approaches focus on employing a flexible, stochastic measure of node similarity based on random walks, *e.g.*, DeepWalk [22], node2vec [2], LINE [26], and HARP [9]. There are several limitations in matrix factorization-based and random walk-based graph learning approaches. First, the embedding function which maps to a low-dimensional vector space is linear or overly simple so that complex patterns cannot be captured; Second, they typically do not incorporate node features; Finally, they are inherently transductive, for the whole graph structure is required in the training phase.

Recently these limitations in graph learning have been addressed by adopting new advances in deep learning. Deep learning with neural networks can represent complex mapping functions and be efficiently optimized by gradient-descent methods. To embed graph nodes to a Euclidean space, deep autoencoders are adopted to extract connectivity patterns from the node similarity matrix or adjacency matrix, *e.g.*, Deep Neural Graph Representations (DNGR) [8] and Structural Deep Network Embeddings (SDNE) [30]. Although autoencoder-based approaches are able to capture more complex patterns than matrix factorization based and random walk based methods, they are still unable to leverage node features.

With celebrated successes of CNN in image recognition, recently, there has been an increased interest in adapting convolutions to graph learning. In [6], the convolution operation is defined in the Fourier domain, that is, the spectral space, of the graph Laplacian. The method is afflicted by two major problems: Firstly, the eigen decomposition is computationally intensive; secondly, filtering in the Fourier domain may result in non-spatially localized effects. In [13], a parameterization of the Fourier filter with smooth coefficients is introduced to make the filter spatially localized. [11] proposes to approximate the filters by using a Chebyshev expansion of the graph Laplacian, which produces spatially localized filters, and also avoids computing the eigenvectors of the Laplacian.

Attention mechanisms have been widely employed in many sequence-based tasks [3][33][16]. Compared with convolution operators, attention mechanisms enjoy two benefits: Firstly, they are able to aggregate any variable sized neighborhood or sequence; further, the weights for aggregation are functions of the contents of a neighborhood or sequence. Therefore, they are adaptive to the contents. [27] adapts an attention mechanism to graph learning and proposes a graph attention network (GAT), achieving current state-of-the-art performance on several graph node classification problems.

## 3. Edge feature enhanced graph neural networks

### 3.1. Architecture overview

Given a graph with $N$ nodes, let $X$ be an $N \times F$ matrix representation of the node features of the whole graph. We denote an element of a matrix or a tensor by indices in the subscript. Specifically, the "$\cdot$" notation in the subscript is used to select the whole range (slice) of a dimension. Therefore, $X_{ij}$ will represent the value of the $j^{th}$ feature of the $i^{th}$ node. $X_{i\cdot} \in \mathbb{R}^F, i = 1, 2, \ldots, N$, represents the $F$-dimensional feature vector of the $i^{th}$ node. Similarly, let $E$ be an $N \times N \times P$ tensor representing the edge features of the graph. Then $E_{ij\cdot} \in \mathbb{R}^P, i = 1, 2, \ldots, N, j = 1, 2, \ldots, N$, represents the $P$-dimensional feature vector of the edge connecting the $i^{th}$ and $j^{th}$ nodes, and $E_{ijp}$ denotes the $p^{th}$ channel of the edge feature in $E_{ij\cdot}$. We use the notation $E_{ij\cdot} = \mathbf{0}$ to mean that there is no edge between the $i^{th}$ and $j^{th}$ nodes. Let $\mathcal{N}_i, i = 1, 2, \ldots, N$, denote the index set of neighboring nodes of node $i$.

Our proposed network has a multi-layer feed-forward architecture. We use superscript $l$ to denote the output of the $l^{th}$ layer. The inputs to the network are denoted by $X^0$ and $E^0$. After passing through the first EGNN layer, $X^0$ is filtered to produce an $N \times F^1$ new node feature matrix $X^1$. In the mean time, edge features are adapted to $E^1$ that preserves the dimensionality of $E^0$. The adapted $E^1$ is fed to the next layer as edge features. This procedure is repeated for every subsequent layer. Within each hidden layer, non-

linear activations can be applied to the filtered node features $X^l$. The node features $X^L$ can be considered as an embedding of the graph nodes in an $F^L$-dimensional space. For a node classification problem, a soft-max operator will be applied to each node embedding vector $X^L_{i\cdot}$ along the last dimension. For a whole-graph prediction (classification or regression) problem, a pooling layer is applied to the first dimension of $X^L$ so that the feature matrix is reduced to a single vector embedding for the whole graph. Then a fully connected layer is applied to the vector, whose output could be used as predictions for regression, or logits for classification. The weights of the network will be trained with supervision from ground truth labels. Figure 1 gives a schematic illustration of the EGNN architecture with a comparison to the existing GNN architectures. Note that the input edge features in $E^0$ are already pre-normalized. The normalization method will be described in the next subsection. Two types of EGNN layers, attention based EGNN (EGNN(A)) layer and convolution based EGNN (EGNN(C)) layer will also be presented in the following subsections.

### 3.2. Doubly stochastic normalization of edges

In graph convolution operations, the edge feature matrices will be used as filters to multiply the node feature matrix. To avoid increasing the scale of output features by multiplication, the edge features need to be normalized. Let $\hat{E}$ be the raw edge features, our normalized features $E$ is produced as follows:

$$\tilde{E}_{ijp} = \frac{\hat{E}_{ijp}}{\sum_{k=1}^{N} \hat{E}_{ikp}} \tag{1}$$

$$E_{ijp} = \sum_{k=1}^{N} \frac{\tilde{E}_{ikp}\tilde{E}_{jkp}}{\sum_{v=1}^{N} \tilde{E}_{vkp}} \tag{2}$$

Note that all elements in $\hat{E}$ are nonnegative. It can be easily verified that such kind of normalized edge feature tensor $E$ satisfies the following properties:

$$E_{ijp} \geq 0, \tag{3}$$

$$\sum_{i=1}^{N} E_{ijp} = \sum_{j=1}^{N} E_{ijp} = 1. \tag{4}$$

In other words, the edge feature matrices $E_{\cdot\cdot p}$ for $p = 1, 2, \cdots, P$ are square nonnegative real matrices with rows and columns summing to 1. Thus, they are doubly stochastic matrices, *i.e.*, they are both left stochastic and right stochastic. Doubly stochastic matrices (DSMs) have several nice properties, *e.g.*, they are symmetric, positive semi-definite and having the largest eigen-value 1. The graph convolution has an effect similar to passing information through edges in a diffusion process by iteratively multiplying the previous result with the edge matrix. Since taking the power of a DSM preserves the three mentioned properties, it prevents the edge matrix from exploding or shrinking to zero during diffusion, thus can help stabilize the process, compared with the previously used row normalization as in GAT [27]:

$$E_{ijp} = \frac{\hat{E}_{ijp}}{\sum_{j=1}^{N} \hat{E}_{ijp}} \tag{5}$$

or symmetric normalization as in GCN [18]:

$$E_{ijp} = \frac{\hat{E}_{ijp}}{\sqrt{\sum_{i=1}^{N} \hat{E}_{ijp}}\sqrt{\sum_{j=1}^{N} \hat{E}_{ijp}}} \tag{6}$$

Further, the powering (or diffusion) has an effect of increasing the gaps between large egien-values, which denoises the edges. The effectiveness of doubly stochastic matrix has been recently demonstrated for graph edges denoising [29].

### 3.3. EGNN(A): Attention based EGNN layer

We describe the attention based EGNN layer. The original GAT model [27] is only able to handle one dimensional binary edge features, *i.e.*, the attention mechanism is defined on the node features of the neighborhood, which does not take the real valued edge features, *e.g.*, weights, into account. To exploit multi-dimensional nonnegative-valued edge features, we propose a new attention mechanism. In our new mechanism, feature vector $X^l_{i\cdot}$ will be aggregated from the feature vectors of the neighboring nodes of the $i^{th}$ node, *i.e.*, $\{X_j, j \in \mathcal{N}_i\}$, by simultaneously incorporating the corresponding edge features. Utilizing the tensor and the notation that zero valued edge features mean no edge connections, the aggregation operation is defined as follows:

$$X^l = \sigma\left[\Big\|_{p=1}^{P} \left(\alpha^l_{\cdot\cdot p}(X^{l-1}, E^{l-1}_{\cdot\cdot p}) g^l(X^{l-1})\right)\right]. \tag{7}$$

Here, $\|$ is the concatenation operator; $\sigma$ is a non-linear activation; $\alpha$ is a function which produces an $N \times N \times P$ tensor; $g$ is a transformation which maps the node features from the input space to the output space, and usually a linear mapping is used:

$$g^l(X^{l-1}) = X^{l-1}W^l, \tag{8}$$

where $W^l$ is an $F^{l-1} \times F^l$ weight matrix.

In Eq. (7), $\alpha^l$ contains the attention coefficients, whose specific entry $\alpha^l_{ijp}$ is a function of $X^{l-1}_{i\cdot}$, $X^{l-1}_{j\cdot}$ and $E_{ijp}$. In existing attention mechanisms [27], the attention coefficient depends on $X_{i\cdot}$ and $X_{j\cdot}$ only. Our mechanism allows the attention operation to be guided by edge features. For multiple dimensional edge features, we consider them as multi-channel signals, and each channel will guide a separate attention operation. The results from different channels

are combined by the concatenation operation. For a specific channel of edge features, our attention function is chosen to be the following:

$$\hat{\alpha}^l_{ijp} = f^l(X^{l-1}_{i\cdot}, X^{l-1}_{j\cdot})E^{l-1}_{ijp}, \tag{9}$$

$$\alpha^l_{\cdot\cdot p} = \mathrm{DS}(\hat{\alpha}^l_{\cdot\cdot p}), \tag{10}$$

where DS is the doubly stochastic normalization operator defined in Eqs. (1) and (2). In principle, $f^l$ can be any ordinary attention function which produces a scalar value from two input vectors. In this paper, we use a linear function as the attention function for simplicity:

$$f^l(X^{l-1}_{i\cdot}, X^{l-1}_{j\cdot}) = \exp\left\{\mathrm{L}\left(a^T[X^{l-1}_{i\cdot}W^l \| X^{l-1}_{j\cdot}W^l]\right)\right\}, \tag{11}$$

where L is the LeakyReLU activation function; $W^l$ is the same mapping as in (8).

The attention coefficients will be used as new edge features for the next layer:

$$E^l = \alpha^l. \tag{12}$$

By doing so, EGNN adapts the edge features across the network layers, which helps capture essential edge features as determined by our novel attention mechanism.

### 3.4. EGNN(C): Convolution based EGNN layer

By regarding the graph convolution operation as a special case of the graph attention operation, we derive our EGNN(C) layer from the formula of EGNN(A) layer. Indeed, the essential difference between GCN [18] and GAT [27] is whether we use the attention coefficients (*i.e.*, matrix $\alpha$) or the adjacency matrix to aggregate node features. With this view, we derive EGNN(C) by replacing the attention coefficient matrices $\alpha_{\cdot\cdot p}$ with the corresponding edge feature matrices $E_{\cdot\cdot p}$. The resulting formula for EGNN(C) is given as follows:

$$X^l = \sigma\left[\Big\|_{p=1}^{P}\left(E_{\cdot\cdot p}X^{l-1}W^l\right)\right], \tag{13}$$

where the notations are the same as in Section 3.3.

### 3.5. Edge features for directed graph

In real world, many graphs are directed. Often times, edge direction contains important information about the graph. For example, in a citation network, machine learning papers sometimes cite mathematics papers or other theoretical papers. However, mathematics papers may seldom cite machine learning papers. In many previous studies including GCNs and GAT, edge directions are not considered. In their experiments, directed graphs such as citation networks are simply treated as undirected graphs. In this paper, we show in the experiment part that discarding edge directions will lose important information. By viewing directions of edges as a kind of edge features, we encode a directed edge channel $E_{ijp}$ to be

$$\left[\hat{E}_{ijp} \quad \hat{E}_{jip} \quad \hat{E}_{ijp} + \hat{E}_{jip}\right].$$

Therefore, each directed channel is augmented to three channels. Note that the three channels define three types of neighborhoods: forward, backward, and undirected. As a result, EGNN will aggregate node information from these three different types of neighborhoods, which contains the direction information. Taking the citation network for instance, EGNN will apply the attention mechanism or convolution operation on the papers that a specific paper cited, the papers cited this paper, and the union of the former two. With such edge features, discriminative patterns in various types of neighborhoods can be effectively captured.

## 4. Experimental results

For all the experiments, we implement the algorithms in Python on the TensorFlow platform [1]. In all the experiments, models are trained with a Nvidia Tesla K40 graphics card with 12 Gigabyte graphics memory.

### 4.1. Citation networks

To benchmark the effectiveness of our proposed models, we apply them to the network node classification problem. Three datasets are tested: Cora [23], Citeseer [23], and Pubmed [20]. Some basic statistics about these datasets are listed in Table 1. All three datasets are directed graphs,

Table 1: Summary of citation network datasets

| | Cora | Citeseer | Pubmed |
|---|---|---|---|
| # Nodes | 2708 | 3327 | 19717 |
| # Edges | 5429 | 4732 | 44338 |
| # Node Features | 1433 | 3703 | 500 |
| # Classes | 7 | 6 | 3 |

where edge directions represent the directions of citations. For Cora and Citeseer, node features contain binary indicators representing the occurrences of predefined keywords in a paper. For Pubmed, term frequency-inverse document frequency (TF-IDF) features are employed to describe the network nodes (*i.e.*, papers).

The three citation network datasets are also used in [32] [18] [27]. However, they all use a pre-processed version which discards the edge directions. Since our EGNN models require the edge directions to construct edge features, we use the original version from [23] and [20]. For each of the

Table 2: Classification accuracies on citation networks. Doubly stochastic normalization, multi-dimensional edge features, edge adaptiveness and weighted loss components are denoted by "D", "M", "A" and "W", respectively, in square brackets.

| Dataset | Cora | | CiteSeer | | Pubmed | |
|---|---|---|---|---|---|---|
| Splitting | Sparse | Dense | Sparse | Dense | Sparse | Dense |
| GCN | $72.9 \pm 0.8\%$ | $72.0 \pm 1.2\%$ | $69.2 \pm 0.7\%$ | $75.3 \pm 0.4\%$ | $83.3 \pm 0.4\%$ | $83.4 \pm 0.2\%$ |
| GAT | $75.5 \pm 1.1\%$ | $79.0 \pm 1.0\%$ | $69.5 \pm 0.5\%$ | $74.9 \pm 0.5\%$ | $83.4 \pm 0.1\%$ | $83.4 \pm 0.2\%$ |
| EGNN(C)[W] | $82.7 \pm 0.6\%$ | $87.6 \pm 0.6\%$ | $69.3 \pm 0.6\%$ | $76.0 \pm 0.5\%$ | $84.5 \pm 0.2\%$ | $84.3 \pm 0.4\%$ |
| EGNN(A)[W] | $82.7 \pm 0.6\%$ | $86.6 \pm 0.6\%$ | $69.4 \pm 0.5\%$ | $74.9 \pm 0.8\%$ | $83.1 \pm 0.2\%$ | $82.7 \pm 0.2\%$ |
| EGNN(C)[D] | $81.8 \pm 0.5\%$ | $85.1 \pm 0.5\%$ | $\mathbf{70.6} \pm 0.3\%$ | $75.0 \pm 0.3\%$ | $84.3 \pm 0.1\%$ | $84.1 \pm 0.1\%$ |
| EGNN(C)[DW] | $83.2 \pm 0.3\%$ | $87.4 \pm 0.4\%$ | $70.3 \pm 0.3\%$ | $75.4 \pm 0.5\%$ | $84.1 \pm 0.1\%$ | $84.1 \pm 0.1\%$ |
| EGNN(C)[M] | $80.2 \pm 0.4\%$ | $86.1 \pm 0.5\%$ | $69.4 \pm 0.3\%$ | $76.8 \pm 0.4\%$ | $\mathbf{86.2} \pm 0.2\%$ | $\mathbf{86.7} \pm 0.1\%$ |
| EGNN(C)[MW] | $82.3 \pm 0.4\%$ | $87.2 \pm 0.4\%$ | $69.4 \pm 0.3\%$ | $\mathbf{77.1} \pm 0.4\%$ | $\mathbf{86.2} \pm 0.1\%$ | $86.4 \pm 0.3\%$ |
| EGNN(C)[DM] | $83.0 \pm 0.3\%$ | $\mathbf{88.8} \pm 0.3\%$ | $69.5 \pm 0.3\%$ | $76.7 \pm 0.4\%$ | $86.0 \pm 0.1\%$ | $86.0 \pm 0.1\%$ |
| EGNN(C)[DMW] | $\mathbf{83.4} \pm 0.3\%$ | $88.5 \pm 0.4\%$ | $69.5 \pm 0.3\%$ | $76.6 \pm 0.4\%$ | $85.8 \pm 0.1\%$ | $85.6 \pm 0.2\%$ |
| EGNN(A)[A] | $76.0 \pm 1.0\%$ | $79.1 \pm 1.0\%$ | $69.5 \pm 0.4\%$ | $74.6 \pm 0.3\%$ | $83.4 \pm 0.1\%$ | $83.6 \pm 0.2\%$ |
| EGNN(A)[AW] | $82.6 \pm 0.6\%$ | $86.3 \pm 0.9\%$ | $69.4 \pm 0.4\%$ | $74.9 \pm 0.4\%$ | $83.7 \pm 0.2\%$ | $82.8 \pm 0.3\%$ |
| EGNN(A)[D] | $80.1 \pm 1.0\%$ | $85.4 \pm 0.5\%$ | $70.1 \pm 0.4\%$ | $74.7 \pm 0.4\%$ | $84.3 \pm 0.2\%$ | $84.2 \pm 0.1\%$ |
| EGNN(A)[DW] | $82.7 \pm 0.4\%$ | $87.2 \pm 0.5\%$ | $69.5 \pm 0.3\%$ | $74.5 \pm 0.5\%$ | $83.9 \pm 0.2\%$ | $83.3 \pm 0.2\%$ |
| EGNN(A)[M] | $81.7 \pm 0.4\%$ | $87.9 \pm 0.4\%$ | $69.4 \pm 0.3\%$ | $75.7 \pm 0.3\%$ | $85.5 \pm 0.1\%$ | $86.0 \pm 0.1\%$ |
| EGNN(A)[MW] | $82.8 \pm 0.3\%$ | $87.0 \pm 0.6\%$ | $69.1 \pm 0.3\%$ | $76.3 \pm 0.5\%$ | $85.2 \pm 0.2\%$ | $85.3 \pm 0.3\%$ |
| EGNN(A)[ADM] | $82.5 \pm 0.3\%$ | $88.4 \pm 0.3\%$ | $69.4 \pm 0.4\%$ | $76.5 \pm 0.3\%$ | $85.7 \pm 0.1\%$ | $\mathbf{86.7} \pm 0.1\%$ |
| EGNN(A)[ADMW] | $83.1 \pm 0.4\%$ | $88.4 \pm 0.3\%$ | $69.3 \pm 0.3\%$ | $76.3 \pm 0.5\%$ | $85.6 \pm 0.2\%$ | $85.7 \pm 0.2\%$ |

three datasets, we split nodes into 3 subsets for training, validation and testing. Two splittings are tested. One splitting has $5\%$, $15\%$ and $80\%$ sized subsets for training, validation and test, respectively. Since it has a small training set, we call it "sparse" splitting. Another splitting has $60\%$, $20\%$ and $20\%$ sized subsets, which is called "dense" splitting.

Following the experiment settings of [18][27], we use two layers of EGNN in all of our experiments for fair comparison. Throughout the experiments in this subsection, we use the Adam optimizer [17] with learning rate $0.005$. An early stopping strategy with a window size of $100$ is adopted for the citation networks; *i.e.*, we stop training if the validation loss does not decrease for $100$ consecutive epochs. We set the output dimension of $W$ to $64$ for hidden layers. We apply dropout [25] with dropout rate $0.6$ to both input features and normalized attention coefficients. $L_2$ regularization with weight decay $0.0005$ is applied to the weights $W$ and $a$. Moreover, exponential linear unit (ELU) [10] is employed as nonlinear activations for hidden layers.

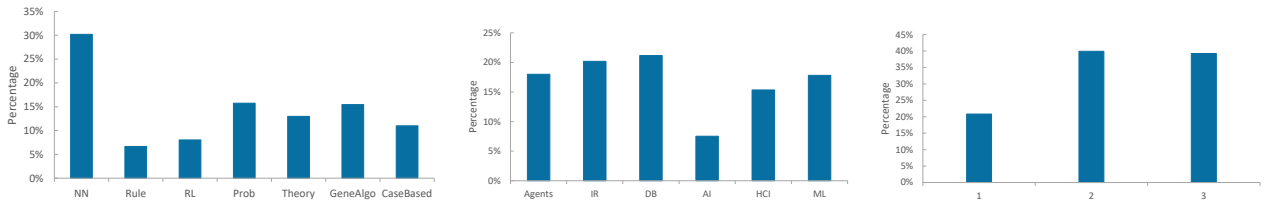
We notice that the class distributions of the training subsets of the three datasets are not balanced. To test the effects of dataset imbalance, we train each algorithm with two different loss functions, *i.e.*, unweighted and weighted losses. The weight of a node belonging to class $k$ is calculated as

$$\frac{\sum_{k=1}^{K} n_k}{K n_k}, \tag{14}$$

where $K$ and $n_k$ are the numbers of classes and nodes belonging to the $k^{th}$ class in the training subset, respectively. Thus, nodes in a minority class are given larger weights and are penalized more in the loss than a majority class.

The baseline methods we used are GCN [18] and GAT [27]. To investigate the effectivenesses of each components, we perform ablation study of EGNN(A) and EGNN(C). We denote a specific version of EGNN(A) or EGNN(C) by EGNN(A)[$\cdot$] or EGNN(C)[$\cdot$], where the letters in the square bracket represent different combinations of components. We denote doubly stochastic normalization, multi-dimensional edge features, edge adaptiveness, and weighted loss by "D", "M", "A" and "W", respectively. For example, EGNN(C)[M] means the EGNN(C) model with the component of the multi-dimensional edge features only. Totally, 18 models are tested, including the two baselines. We run each model 20 times, and record the mean and standard deviation of the classification accuracies, which are listed in Table 2. We can observe several interesting phenomena, some of which warrant further investigations:

- Overall, almost all EGNN variants outperform their corresponding baselines, which indicates that all the three components are able to incorporate useful information for classification. Particularly, multi-dimensional edge features and doubly stochastic normalization improve more than edge adaptiveness.




(a) Cora (b) Citeseer (c) Pubmed

Figure 2: Node class distribution of the training subsets of the three citation networks. The Cora dataset is more imbalanced than the other two.

- The two baselines underperform on both the sparse and dense splittings of the Cora dataset. This is caused by the class imbalance of the Cora dataset. We illustrate the class distributions of the three datasets in Figure 2, from which we can see that Cora is more imbalanced than Citeseer and Pubmed. On the Cora dataset, EGNN(A)[W] and EGNN(C)[W], which add the weighted loss component only, perform much better than GAT and GCN. This verifies that the underperformance of GAT and GCN are caused by the class imbalance. Our proposed models are highly resilient to class imbalance. Without weighted training, our framework obtains high accuracies on the Cora dataset. Weighted training does not always improve performance, especially on less imbalanced datasets, *e.g.*, Pubmed. This indicates that simply weighting the classes is not sufficient to fully solve the class imbalance problem. More sophisticated methods need to be designed to address this problem in the future.

- Performances on dense splittings are consistently higher than on sparse splitting. It is not unexpected because more training data gives an algorithm more information to tune parameters.

- We noticed that the Pubmed graph is much bigger than Cora and Citeseer. Also, its node features (TF-IDF) are more sophisticated than the bag-of-words features of the other two datasets. Finally, Pubmed is the least imbalanced. These characteristics may be the reason that the sparse and dense splittings of Pubmed get close accuracies as the sparse splitting already provides sufficient information for training.

- Another characteristic we noticed is that the average degree of Citeseer is the smallest among the citation networks. Therefore, a very limited number of edges can be utilized in a sparse splitting. This restricts the ability of our models to incorporate edge features; thus, we do not observe much performance gain on the sparse splitting of Citeseer.

- Either EGNN(C)[DW] or EGNN(C)[MW] is not as good as EGNN(C)[W] on the dense splitting of the Cora dataset. However, EGNN(C)[DMW] is better than EGNN(C)[W]. This interesting phenomenon indicates that doubly stochastic normalization and multi-dimensional edge feature might not work well individually on some datasets, but can improve performance considerably if combined.

To compare the computational efficiency of the models, we record the average training time on the Cora dataset with dense splitting in Table 3. Note that EGNN(C) and EGNN(A) are full models with all components implemented. According to Table 3, the proposed EGNN(C) and EGNN(A) are a little higher but still essentially comparable to GCN and GAT in terms of time complexity.

Table 3: Average training time (in milliseconds) per epoch on the Cora dataset with dense splitting.

| Model | GCN | GAT | EGNN(C) | EGNN(A) |
|---|---|---|---|---|
| Time (ms) | 22 | 49 | 48 | 159 |

### 4.2. Molecular analysis

One promising application of graph learning is molecular analysis. A molecule can be represented as a graph, where each atom is a node and chemical bonds are edges. Unlike citation network analysis in Section 4.1, the problem here is whole-graph prediction, either classification or regression. For example, given a graph representation of a molecule, the goal may be to classify it as toxic or not, or to predict the solubility (regression). In other words, we need to predict one value for the whole graph, rather than one value for a graph node. Usually, for each chemical bond, there are several attributes associated with it, *e.g.*, Atom Pair Type, Bond Order, and Ring Status. Therefore, the graphs intrinsically contain multi-dimensional edge features.

Three datasets, Tox21, Lipophilicity and Freesolv, are used to test our algorithms. Tox21 contains 7831 environmental compounds and drugs. Each compound is associated with 12 labels, *e.g.*, androgen receptor, estrogen receptor, and mitochondrial membrane potential, which defines

Table 4: Performance on molecular datasets

| Dataset | Tox21 (AUC) Validation | Tox21 (AUC) Test | Lipo (RMSE) Validation | Lipo (RMSE) Test | Freesolv (RMSE) Validation | Freesolv (RMSE) Test |
|---|---|---|---|---|---|---|
| RF | $0.78 \pm 0.01$ | $0.75 \pm 0.03$ | $0.87 \pm 0.02$ | $0.86 \pm 0.04$ | $1.98 \pm 0.07$ | $1.62 \pm 0.14$ |
| Weave | $0.79 \pm 0.02$ | $0.80 \pm 0.02$ | $0.88 \pm 0.06$ | $0.89 \pm 0.04$ | $1.35 \pm 0.22$ | $1.37 \pm 0.14$ |
| EGNN(C) | $\mathbf{0.82} \pm 0.01$ | $\mathbf{0.82} \pm 0.01$ | $0.80 \pm 0.02$ | $\mathbf{0.75} \pm 0.01$ | $\mathbf{1.07} \pm 0.08$ | $1.09 \pm 0.08$ |
| EGNN(A) | $\mathbf{0.82} \pm 0.01$ | $0.81 \pm 0.01$ | $\mathbf{0.79} \pm 0.02$ | $\mathbf{0.75} \pm 0.01$ | $1.09 \pm 0.12$ | $\mathbf{1.01} \pm 0.12$ |

a multi-label classification problem. Lipophilicity contains 4200 compounds. The goal is to predict compound solubility, which is a regression task. Freesolv includes a set of 642 neutral molecules, which similarly defines a regression task. For all the three datasets, compounds are converted to graphs. For all the three datasets, nodes are described by 25-dimensional feature vectors. The dimensionality of edge feature vectors are 42, 21, and 25 for Tox21, Lipo, and Freesolv, respectively.

For both EGNN(A) and EGNN(C), we implement a network consisting of 2 graph processing layers, a global max-pooling layer, and a fully connected layer. For each graph processing layer, the output dimensions of the linear mapping $g$ are fixed to be 16. For Tox21, the sigmoid cross entropy loss is applied to the output logits of the fully connected layer. For Lipo and Freesolv, the mean squared error loss is employed. The networks are trained by Adam optimizer [17] with learning rate 0.0005. An early stopping strategy with a window size of 200 is adopted. $L_2$ regularization with weight decay 0.0001 is applied to parameters of the models except for bias parameters. Moreover, exponential linear unit (ELU) [10] is employed as nonlinear activations for hidden layers.

Our models are compared with two baseline models which are shown in MoleculeNet [31]: Random Forest and Weave. Random Forest is a traditional learning model which is widely applied to various problems. Weave model [15] is similar to graph convolution but specifically designed for molecular analysis.

All the three datasets are split into training, validation and test subsets with a ratio of 8:1:1. We run our models 5 times, and record the means and standard deviations of performance scores. For classification task (*i.e.*, Tox21), Area Under Curve (AUC) scores of the receiver operating characteristic (ROC) curve are recorded. Since it is a multi-label classification problem, we record the AUCs of each class and take the averaged value as the final score. For regression (*i.e.*, Lipo and Freesolv), root mean square error (RMSE) is used as the evaluation metric. The scores are given in Table 4. The results show that EGNN(C) and EGNN(A) outperform the two baselines with considerable margins. On the Tox21 dataset, the AUC scores are improved by more than 0.2 compared with the Weave model. For the two regression tasks, RMSEs are improved by about 0.1 and 0.3 on the Lipo and Freesolv datasets, respectively. The scores of EGNN(C) and EGNN(A) are close on the three datasets.

## 5. Conclusions

In this paper, we propose a new framework to address the existing problems in the current state-of-the-art graph neural network models. Specifically, we propose a new attention mechanism to incorporate multi-dimensional nonnegative-valued edge features. Then, we propose a new graph neural network architecture that adapts edge features across neural network layers. Our framework admits a formula that allows for extending convolutions to multi-dimensional edge features. Further, we propose to use doubly stochastic normalization, as opposed to the ordinary row normalization or symmetric normalization used in the existing graph neural network models. Finally, we propose a method to design multi-dimensional edge features for directed edges to effectively handle directed graphs. Extensive experiments are conducted on three citation network datasets for graph node classification evaluation, and on three molecular datasets to test the performance on whole graph classification and regression tasks. Experimental results show that our new framework outperforms current state-of-the-art models such as GCN and GAT significantly and consistently on all the datasets. Detailed ablation study also shows the effectiveness of each individual component in our framework.

## References


[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A System for Large-scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[2] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola. Distributed Large-scale Natural

Graph Factorization. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 37–48, New York, NY, USA, 2013. ACM.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*, Sept. 2014.

[4] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, page 7, 2001.

[5] S. Bhagat, G. Cormode, and S. Muthukrishnan. Node Classification in Social Networks. *arXiv:1101.3291 [physics]*, pages 115–148, 2011.

[6] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral Networks and Locally Connected Networks on Graphs. *arXiv:1312.6203 [cs]*, Dec. 2013.

[7] S. Cao, W. Lu, and Q. Xu. GraRep: Learning Graph Representations with Global Structural Information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 891–900, New York, NY, USA, 2015. ACM.

[8] S. Cao, W. Lu, and Q. Xu. Deep Neural Networks for Learning Graph Representations. In *AAAI Conference on Artificial Intelligence*, 2016.

[9] H. Chen, B. Perozzi, Y. Hu, and S. Skiena. HARP: Hierarchical Representation Learning for Networks. In *AAAI Conference on Artificial Intelligence*, 2018.

[10] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In *International Conference on Learning Representations*, 2016.

[11] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 3844–3852. Curran Associates, Inc., 2016.

[12] J. L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, Mar. 1990.

[13] M. Henaff, J. Bruna, and Y. LeCun. Deep Convolutional Networks on Graph-Structured Data. *arXiv:1506.05163 [cs]*, June 2015.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735, Nov. 1997.

[15] S. Kearnes, K. Mccloskey, M. Berndl, V. Pande, and P. Riley. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer - Aided Molecular Design; Dordrecht*, 30(8):595–608, Aug. 2016.

[16] S. Kim, J.-H. Hong, I. Kang, and N. Kwak. Semantic Sentence Matching with Densely-connected Recurrent and Co-attentive Information. *arXiv:1805.11360 [cs]*, May 2018.

[17] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

[18] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*, 2017.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov. 1998.

[20] G. Namata, B. London, L. Getoor, and B. Huang. Query-driven Active Surveying for Collective Classification. In *Workshop on Mining and Learning with Graphs*, 2012.

[21] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric Transitivity Preserving Graph Embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1105–1114, New York, NY, USA, 2016. ACM.

[22] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online Learning of Social Representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM.

[23] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective Classification in Network Data. *AI Magazine; La Canada*, 29(3):93–106, 2008.

[24] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-Lehman Graph Kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[26] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. LINE: Large-scale Information Network Embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, 2015.

[27] P. Velickovic, G. Cucurull, A. Casanova, and A. Romero. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

[28] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph Kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

[29] B. Wang, A. Pourshafeie, M. Zitnik, J. Zhu, C. D. Bustamante, S. Batzoglou, and J. Leskovec. Network Enhancement: A general method to denoise weighted biological networks. *arXiv:1805.03327 [cs, q-bio]*, May 2018.

[30] D. Wang, P. Cui, and W. Zhu. Structural Deep Network Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1225–1234, New York, NY, USA, 2016. ACM.

[31] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

[32] Z. Yang, W. Cohen, and R. Salakhudinov. Revisiting Semi-Supervised Learning with Graph Embeddings. In *International Conference on Machine Learning*, pages 40–48, June 2016.

[33] G. Zhou, C. Song, X. Zhu, X. Ma, Y. Yan, X. Dai, H. Zhu, J. Jin, H. Li, and K. Gai. Deep Interest Network for Click-Through Rate Prediction. *arXiv:1706.06978 [cs, stat]*, June 2017.