

# Analyzing Performance Bottleneck in Graph Neural Network Training: An Experimental View

Zhaokang Wang, Yunpan Wang, Chunfeng Yuan, Yihua Huang\*

*State Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China*

---

## Abstract

This is the abstract.

*Keywords:* Keyword 1

---

## 1. Introduction

This is a survey [? ].

## 2. Review of Graph Neural Networks

In this section, we introduce the concepts related to the graph neural network (GNN, for short) and briefly survey typical graph neural networks. We denote a simple graph  $\mathcal{G}$  as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  and  $\mathcal{E}$  are the vertex set and the edge set of  $\mathcal{G}$ , respectively. Let  $n = |\mathcal{V}|$  and  $m = |\mathcal{E}|$  as the number of vertices/edges. We use  $v_i$  ( $0 \leq i < n$ ) to denote a vertex and  $e_{i,j} = (v_i, v_j)$  to denote the edge pointing from  $v_i$  to  $v_j$ . The adjacency set of  $v_i$  is  $\mathcal{N}(v_i) = \{v | (v_i, v) \in \mathcal{E}\}$ . We denote a *vector* with a bold lower case letter like  $\mathbf{x}$  and a *matrix* with a bold upper case letter like  $\mathbf{X}$ .

### 2.1. General Structure of Graph Neural Networks

As illustrated in Figure 1, a typical GNN can be decomposed into three parts: an input layer + several GNN layers + a prediction layer.

---

\*Corresponding author

Email address: {wangzhaokang, wangyp}@smail.nju.edu.cn, {cfyuan, yhuang}@nju.edu.cn (Zhaokang Wang, Yunpan Wang, Chunfeng Yuan, Yihua Huang)

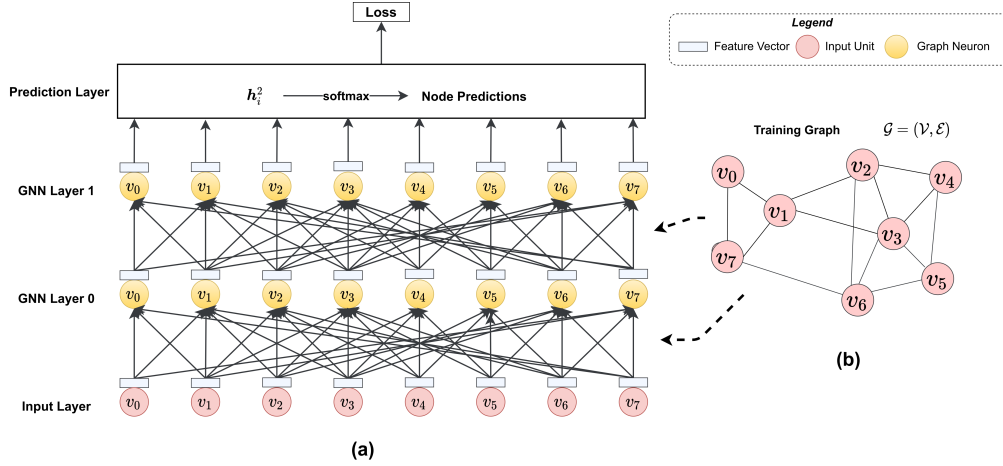


Figure 1: Structure of a typical graph neural network. (a) Demo GNN, (b) Demo graph. The target application is the node classification. The demo GNN has two GNN layers.

15 A GNN receives a graph  $\mathcal{G}$  as the input. Every vertex  $v_i$  in  $\mathcal{G}$  is attached with a feature vector  $\mathbf{x}_i$  to describe the properties of the vertex. The edges of  $\mathcal{G}$  may also be attached with feature vectors  $\mathbf{e}_{i,j}$ . The input layer of a GNN receives feature vectors from all vertices and passes them to GNN layers.

A GNN layer consists of  $n$  graph neurons, where  $n$  is the number of vertices  
 20 in  $\mathcal{G}$ . Each graph neuron corresponds to a vertex in  $\mathcal{G}$ . In the first GNN layer (Layer 0), the graph neuron of the vertex  $v_i$  collects input feature vectors of itself and the vertices  $\mathbf{x}_j$  that are adjacent to  $v_i$  in  $\mathcal{G}$  (i.e.,  $v_j \in \mathcal{N}(v_i)$ ) from the input layer. After aggregating input feature vectors and applying non-linear transformation, the graph neuron outputs a hidden feature vector  $\mathbf{h}_i^1$   
 25 for  $v_i$ . Take the demo DNN in Figure 1(a) as the example. Since  $\mathcal{N}(v_3) = \{v_1, v_2, v_4, v_5, v_6\}$ , the graph neuron of  $v_1$  at layer 0 collects the feature vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$  from the input layer and outputs  $\mathbf{h}_1^1$ . Different GNNs mainly differ in the graph neurons that they use. We elaborate on their details later.

30 The connection between the input layer and the first GNN layer is determined by the topology of  $\mathcal{G}$ . In the traditional neural networks, neurons of neighboring layers are fully connected. In GNNs, two graph neurons are con-

nected only if their corresponding vertices have an edge between them in  $\mathcal{G}$ . Most real-world graphs are very *sparse*, i.e.  $|\mathcal{E}| \ll |\mathcal{V}|^2$ .

35 In the next GNN layer (Layer 1), the graph neuron of  $v_i$  collects the hidden feature vectors of itself  $\mathbf{h}_i^1$  and its neighbors ( $\mathbf{h}_j^1$  with  $v_j \in \mathcal{N}(v_i)$ ) from the *previous* GNN layer. Based on the collected hidden vectors, the graph neuron in Layer 1 outputs a new hidden feature vector  $\mathbf{h}_i^2$  for  $v_i$ . Though there are only two GNN layers in Figure 1, a GNN allows to stack more GNN layers to  
40 support deeper graph analysis.

Assume there are  $L$  GNN layers. The last GNN layer (Layer  $L - 1$ ) outputs a hidden feature vector  $\mathbf{h}_i^L$  for every vertex  $v_i$ . As an embedding vector,  $\mathbf{h}_i^L$  encodes the knowledge learned from the input layer and all the previous GNN layers. Since  $\mathbf{h}_i^L$  is affected by  $v_i$  and the vertices in the  $L$ -hop neighborhood of  
45  $v_i$ , analyzing a graph with a *deeper* GNN means analyzing each vertex with a *wider* scope.

The hidden feature vectors  $\mathbf{h}_i^L$  of the last GNN layer are fed to the prediction layer to generate the output of the whole GNN. The prediction layer is a standard neural network. The structure of the prediction layer depends on the  
50 prediction task of the GNN. Take the node classification task as the example, as shown in Figure 1. The node classification predicts a label for every vertex in  $\mathcal{G}$ . In this case, the prediction layer can be a simple softmax layer with  $\mathbf{h}_i^L$  as the input and a vector of probabilities as the output. If the prediction task is edge prediction, the hidden feature vectors of two vertices are concatenated and  
55 fed into a softmax layer. If we need to predict a label for the whole graph, a pooling (max/mean/...) layer is added to generate an embedding vector for the whole graph and the embedding vector is used to produce the final prediction.

Supporting end-to-end training is a prominent advantage of GNN, compared with other graph-based machine learning methods. We can calculate the gra-  
60 dients of the loss function on the model parameters from the prediction layer directly. With the help of the back propagation technique, the gradient is propagated from the prediction layer back to the previous GNN layers layer by layer. The model parameters are updated with a gradient descent optimizer like

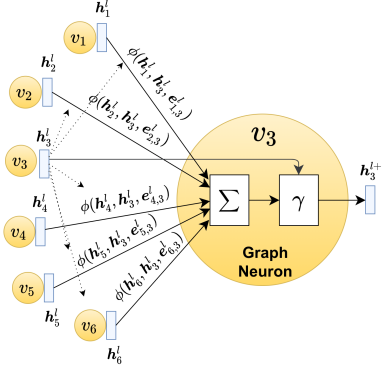


Figure 2: Graph neuron of  $v_3$  at the GNN layer  $l$  with the graph  $\mathcal{G}$  in Figure 1(b).  $\phi/\Sigma/\gamma$  are the message/aggregation/vertex update functions in the message-passing model, respectively.

Adam. Except for the input feature vector, there is no need to conduct hand-  
65 worked feature extraction. In a fully parameterized way, the GNN automatically  
extracts an embedding vector for each vertex from its  $L$ -hop neighborhood. The  
parameters are tuned according to the specific prediction task, leading to high  
prediction accuracy.

## 2.2. Graph Neuron and Message-passing Model

70 Graph neurons are building blocks of a GNN. A GNN layer consists of  $|\mathcal{V}|$   
graph neurons. Each vertex corresponds to a graph neuron. A graph neuron  
as shown in Figure 2 is a small neural network. The graph neuron of  $v_i$  at  
layer  $l$  receives hidden feature vectors  $\mathbf{h}_j^l$  from the graph neurons of  $v_i$  and its  
neighbors ( $v_j \in \{v_i\} \cup \mathcal{N}(v_i)$ ) at the previous GNN layer <sup>1</sup>. The graph neuron  
75 aggregates the received hidden feature vectors, applies non-linear transforma-  
tions, and outputs a new hidden feature vector  $\mathbf{h}_i^{l+1}$ .

We follow the message-passing model [?] to formally define a graph neuron.

<sup>1</sup>For the GNN layer 0, graph neurons receive input feature vectors, i.e.,  $\mathbf{h}_i^0 = \mathbf{x}_i$

The message-passing model is widely used in the cutting-edge GNN training systems like PyTorch Geometric (PyG) [?] and Deep Graph Library (DGL) [?]. Figure 2 shows the structure of a graph neuron in the message-passing model. Graph neurons at layer  $l$  are made of three *differentiable* functions:  $\phi^l$ ,  $\Sigma^l$  and  $\gamma^l$ . The graph neuron calculates the output hidden vector  $\mathbf{h}_i^{l+1}$  by

$$\mathbf{h}_i^{l+1} = \gamma^l(\mathbf{h}_i^l, \Sigma_{v_j \in \mathcal{N}(v_i)}^l \phi^l(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{e}_{j,i})).$$

$\phi^l$  is the *message* function. For every incident edge  $(v_j, v_i)$  of  $v_i$ ,  $\phi$  receives the output hidden feature vectors  $\mathbf{h}_i^l$  and  $\mathbf{h}_j^l$  of the previous GNN layer and the edge feature vector  $\mathbf{e}_{j,i}$  as the input.  $\phi^l$  outputs a message vector  $\mathbf{m}_{j,i}^l$  for every edge  $(v_j, v_i)$  at layer  $l$ , i.e.,  $\mathbf{m}_{j,i}^l = \phi^l(\mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{e}_{j,i})$ . For  $v_i$ , the message vectors  $\mathbf{m}_{x,j}^l$  with  $v_x \in \mathcal{N}(v_i)$  are aggregated by the *aggregation* function  $\Sigma^l$  to produce an aggregated vector  $\mathbf{s}_i^l$ , i.e.,  $\mathbf{s}_i^l = \Sigma_{v_j \in \mathcal{N}(v_i)}^l \mathbf{m}_{j,i}^l$ .  $v_i$ 's aggregated vector  $\mathbf{s}_i^l$  and its hidden vector  $\mathbf{h}_i^l$  from the previous GNN layer are fed into the *vertex update* function  $\gamma^l$  to calculate the output hidden vector  $\mathbf{h}_i^{l+1}$  of the current layer  $l$ , i.e.,  $\mathbf{h}_i^{l+1} = \gamma^l(\mathbf{h}_i^l, \mathbf{s}_i^l)$ . The end-to-end training requires  $\phi^l$  and  $\gamma^l$  (like multi layer perceptrons and GRU) and  $\Sigma_l$  (like mean, sum, element-wise min/max) are *differentiable* to make the whole GNN differentialble.

Different GNNs adopt different kinds of graph neurons and have different definitions of the three functions.  $\phi$  and  $\Sigma$  are the *edge computation* functions. They are conducted over every edge in  $\mathcal{G}$ .  $\gamma$  is the *vertex computation* function. It is conducted over every vertex in  $\mathcal{G}$ . Table 1 and Table 2 list the edge functions and the vertex functions of typical GNNs, respectively.

### 2.3. Classification of GNNs

Since we focus on analyzing the performance bottleneck in training GNNs, we classify the typical GNNs from the view of computational complexity. The computational complexity of a GNN layer is related to the complexity of its vertex and edge functions, i.e.  $O(m * (O_\phi + O_\Sigma) + n * O_\gamma)$ , where  $O_\phi/O_\Sigma/O_\gamma$  are the computational complexity of the three functions.

GNN	Type	$\Sigma$	$\phi$	Complexity
ChebNet [? ]	Spectral	sum	$\mathbf{m}_{j,i,,k}^l = T_k(\tilde{L})_{j,i} \mathbf{h}_j^l$	$O(K * d_{in})$
<b>GCN</b>	Spectral	sum	$\mathbf{m}_{j,i}^l = e_{j,i} \mathbf{h}_j^l$	$O(d_{in})$
AGCN	Spectral	sum	$\mathbf{m}_{j,i}^l = \tilde{e}_{j,i}^l \mathbf{h}_j^l$	$O(d_{in})$
GraphSAGE	Non-spectral	mean/LSTM	$\mathbf{m}_{j,i}^l = \mathbf{h}_j^l$	$O(1)$
Neural FPs	Non-spectral	sum	$\mathbf{m}_{j,i}^l = \mathbf{h}_j^l$	$O(d_{in})$
SSE	Recurrent	sum	$\mathbf{m}_{j,i}^l = [\mathbf{h}_i^l \parallel \mathbf{h}_j^l]$	$O(1)$
<b>GGNN</b>	Gated	sum	$\mathbf{m}_{j,i} = \mathbf{W}^l \mathbf{h}_j^l$	$O(d_{in} * d_{out})$
<b>GAT</b>	Attention	sum	$\alpha_{j,i}^k = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}^{l,k} \mathbf{h}_j^l \parallel \mathbf{W}^{l,k} \mathbf{h}_i^l]))}{\sum_{k \in \mathcal{N}(i)} \exp(\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}^{l,k} \mathbf{h}_j^l \parallel \mathbf{W}^{l,k} \mathbf{h}_k^l]))}$ $\mathbf{m}_{j,i}^l = \parallel_{k=1}^K \delta(\alpha_{j,i}^k \mathbf{W}^{l,k} \mathbf{h}_j^l)$	$O(K * d_{in} * d_{out})$
<b>GaAN</b>	Attention	sum,max,mean	$\alpha_{j,i}^k = \frac{\exp(\mathbf{a}^T [\mathbf{W}_{xa}^{l,k} \mathbf{h}_j^l \parallel \mathbf{W}_{ya}^{l,k} \mathbf{h}_i^l])}{\sum_{k \in \mathcal{N}(j)} \exp(\mathbf{a}^T [\mathbf{W}_{xa}^{l,k} \mathbf{h}_j^l \parallel \mathbf{W}_{ya}^{l,k} \mathbf{h}_k^l])}$ $\mathbf{m}_{j,i,1}^l = \parallel_{k=1}^K \delta(\alpha_{j,i}^k \mathbf{W}_v^{l,k} \mathbf{h}_j^l)$ $\mathbf{m}_{j,i,2}^l = \mathbf{W}_m^l \mathbf{h}_j^l$ $\mathbf{m}_{j,i,3}^l = \mathbf{h}_j^l$	$O(\max(d_a, d_m) * K * d_{in})$

Table 1: Typical graph neural networks and their edge computation functions.  $d_{in}$  and  $d_{out}$  are dimensions of the input and output hidden feature vectors, respectively. Blue variables are model parameters to learn.

GNN	$\gamma$	Complexity
ChebNet [? ]	$\mathbf{h}_i^{l+1} = \sum_{k=0}^K \mathbf{W}^{l,k} \mathbf{s}_{i,k}^l$	$O(K * d_{in} * d_{out})$
<b>GCN</b>	$\mathbf{h}_i^{l+1} = \mathbf{W}^l \mathbf{s}_i^l$	$O(d_{in} * d_{out})$
AGCN	$\mathbf{h}_i^{l+1} = \mathbf{W}^l \mathbf{s}_i^l$	$O(d_{in} * d_{out})$
GraphSAGE	$\mathbf{h}_i^{l+1} = \delta(\mathbf{W}^l [\mathbf{s}_i^l \parallel \mathbf{h}_i^l])$	$O(d_{in} * d_{out})$
Neural FPs	$\mathbf{h}_i^{l+1} = \delta(\mathbf{h}_i^l + \mathbf{W}^{N_i} \mathbf{s}_i^l)$	$O(d_{in} * d_{out})$
SSE	$\mathbf{h}_i^{l+1} = (1 - \alpha) \mathbf{h}_i^l + \alpha \delta(\mathbf{W}_1^l \delta(\mathbf{W}_2^l), \mathbf{s}_i^l)$	$O(d_{in} * d_{out})$
<b>GGNN</b>	$\mathbf{z}_i^l = \delta(\mathbf{W}^z \mathbf{s}_i^l + \mathbf{b}^{sz} + \mathbf{U}^z \mathbf{h}_i^l + \mathbf{b}^{hz})$ $\mathbf{r}_i^l = \delta(\mathbf{W}^r \mathbf{s}_i^l + \mathbf{b}^{sr} + \mathbf{U}^r \mathbf{h}_i^l + \mathbf{b}^{hr})$ $\mathbf{h}_i^{l+1} = \tanh(\mathbf{W}^l \mathbf{s}_i^l + \mathbf{b}^s + \mathbf{U}(\mathbf{r}_i^l \odot \mathbf{h}_i^l + \mathbf{b}^h))$ $\mathbf{h}_i^{l+1} = (1 - \mathbf{z}_i^l) \odot \mathbf{h}_i^l + \mathbf{z}_i^l \odot \mathbf{h}_i^{l+1}$	$O(\max(d_{in}, d_{out}) * d_{out})$
<b>GAT</b>	$\mathbf{h}_i^{l+1} = \mathbf{s}_i^l$	$O(1)$
<b>GaAN</b>	$\mathbf{g}_i = \mathbf{W}_g^l [\mathbf{h}_i^l \parallel \mathbf{s}_{i,2}^l \parallel \mathbf{s}_{i,3}^l]$ $\mathbf{h}_i^{l+1} = \mathbf{W}_o^l [\mathbf{h}_i^l \parallel (\mathbf{g}_i \odot \mathbf{s}_{i,3}^l)]$	$O(\max(d_{in} + K * d_v, 2 * d_{in} + d_m) d_{out})$

Table 2: Typical graph neural networks and their vertex computation functions.  $d_{in}$  and  $d_{out}$  are dimensions of the input and output hidden feature vectors, respectively. Blue variables are model parameters to learn.

The complexity can be decomposed into two parts: the edge computation complexity  $O_\phi + O_\Sigma$  and the vertex computation complexity  $O_\gamma$ . In Table 1 and Table 2, we list the edge and vertex computation complexity, respectively. The edge/vertex complexity of a graph neuron are affected by the dimensions of the input/output hidden vectors  $d_{in}$  and  $d_{out}$  and the dimensions of the model parameters (like the number of heads  $K$  in GAT and the dimensions of the view vectors  $d_a/d_v$  in GaAN).

We classify the typical GNNs into four groups based on their edge/vertex complexity.

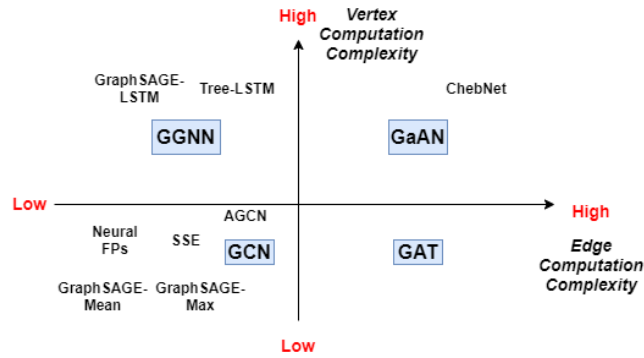


Figure 3: Complexity quadrant of typical GNNs. If two GNNs have the same big-O complexity, we further compare their complexity according to the number of parameters.

### 3. Experiment Design

### 4. Experiment Results and Analysis

### 5. Insights

### 6. Related Work

### 7. Conclusion and Future Work

### Acknowledgment