Response to the Reviewers

We thank the reviewers for their critical assessment and insightful comments of our work. We have made extensive modifications to our manuscript. In the following we address their concerns point by point. We also prepare an *annotated version* for our revised manuscript. In the annotated version, changes corresponding to each point are all highlighted by red squares.

Summary of Changes

Reviewer 1

Reviewer Point P 1.1 — The authors conduct an empirical analysis of performance bottlenecks in graph neural network training. The authors identify the edge-related calculation is the performance bottleneck. Experimental of several GNN variants, such as GCN, GGNNNN, GAT and GaAN on six real-world graph datasets verify the importance of the findings. The experimental analysis is sufficient. However, there are some tiny issues in this paper.

Reply: Thank you for your positive comments on our manuscript. We have carefully revised the manuscript according to your comments. Based on the suggestions, we have made an extensive modification on the revised manuscript. Please see the detailed responses below. The changes corresponding to each issue were highlighted by red squares in the annotated version of our manuscript.

Reviewer Point P 1.2 — There are lots of symbols in this paper. Some symbols are reused and confusing, such as s denotes sub-layers or edge features.

Reply: We apologize for the confusing use of symbols and thank you for pointing the problem out. To clarify the symbol usage, we have checked the manuscript and unified the usage of symbol. We avoid the problem of reusing symbols, so that each symbol only represents one meaning after the revision. We summarize the frequently-used symbols in Table 1 in the revised manuscript. We quote the table below:

Category	Symbol	Meaning
Graph Structure	$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	The simple undirected input graph with the vertex set $\mathcal{V}$ and the edge set $\mathcal{E}$.
	v_x	The x-th vertex of the input graph.
	$e_{y,x}$	The edge pointing from v_y to v_x of the input graph.
	$\mathcal{N}(v_x)$	The adjacency set of v_x in the input graph.
	$\bar{d}$	The average degree of the input graph.
GNN Definition	L	The number of GNN layers.
	K	The number of heads in a GNN layer.
	ϕ^l	The messaging function of the GNN layer l.
	Σ^l	The aggregation function of the GNN layer l.
	γ^l	The vertex updating function of the GNN layer l.
	$\phi^{l,i}$ / $\Sigma^{l,i}$ / $\gamma^{l,i}$	The messaging/aggregation/updating function of the i-th sub-layer of the GNN layer l.
	$\boldsymbol{W}^l, \boldsymbol{W}^{(k)}/\boldsymbol{b}, \boldsymbol{a}$	The matrices/vectors represented by the blue characters are the weight matrices/vectors that need to be learned in the GNN.
Vector	$\boldsymbol{v}_x$	The feature vector of the vertex v_x.
	$\boldsymbol{e}_{y,x}$	The feature vector of the edge $e_{y,x}$.
	$\boldsymbol{h}_x^l$	The input hidden vector of the graph neuron corresponding to v_x in the GNN layer l.
	$\boldsymbol{h}_x^{l+1}$	The output hidden vector of the graph neuron corresponding to v_x in the GNN layer l.
	$\boldsymbol{m}_{y,x}^l$	The message vector of the edge $e_{y,x}$ outputted by ϕ^l of the GNN layer l.
	$\boldsymbol{s}_x^l$	The aggregated vector of the vertex v_x outputted by Σ^l of the GNN layer l.
	$\boldsymbol{h}_x^{l,i}$ / $\boldsymbol{m}_{y,x}^{l,i}$ / $\boldsymbol{s}_x^{l,i}$	The hidden/message/aggregated vector of the vertex v_x outputted by $\gamma^{l,i}$/$\phi^{l,i}$/$\Sigma^{l,i}$ of the i-th sub-layer of the GNN layer l.
	d_{in}^l, d_{out}^l	The dimension of the input/output hidden vectors of the GNN layer l.
	$dim(\boldsymbol{x})$	The dimension of a vector $\boldsymbol{x}$.

In the revised manuscript, we use $e_{y,x}$ to represent an edge and use $\boldsymbol{e}_{y,x}$ to represent its input feature vector. The input feature vectors of all edges are same for all GNN layers. We use $\boldsymbol{s}$ to represent aggregated vectors outputted by the aggregation function Σ in graph neurons. For every vertex v_x, we use $\boldsymbol{s}_x^l$ to denote its aggregated vector in the GNN layer l. If the GNN layer l has sub-layers, $\boldsymbol{s}_x^{l,i}$ represents its aggregated vector in the i-th sub-layer.

To clarify the concept of *sub-layers* in a GNN layer, we have added more description on it in the revised manuscript. We first introduce the concept of sub-layer in Section 2.2 "Graph Neuron and Message-passing Model" as:

> **Section 2.2**
> Some complex GNNs like GAT [6] and GaAN [7] use more than one message passing phase in each GNN layer. We regard every message passing phase in a GNN layer as a *sub-layer*. We will give out more details on sub-layers when we introduce GAT.

We then use GAT as an example to elaborate on the concept of sub-layers in Section 2.3 "Representative GNNs" as:

Section 2.3

Each GAT layer consists of a vertex pre-processing phase and two sub-layers (i.e., message-passing phases).

The vertex pre-processing phase calculates the attention vector $\hat{\boldsymbol{h}}_x^l$ for every vertex v_x by $\hat{\boldsymbol{h}}_x = \|_{k=1}^K \boldsymbol{W}_{(k)}^l \boldsymbol{h}_x^l$. We denote the attention sub-vector generated by the k-th head as $\hat{\boldsymbol{h}}_x[k] = \boldsymbol{W}_{(k)}^l \boldsymbol{h}_x^l$.

The first sub-layer of GAT (defined in Equation 1) uses the attention vectors to emit the attention weight vector $\boldsymbol{m}_{y,x}^{l,0}$ for every edge $e_{y,x}$ and aggregates the attention weight vectors for every vertex v_x to get the weight sum vector $\boldsymbol{h}_x^{l,0}$.

$$\begin{aligned} \boldsymbol{m}_{y,x}^{l,0} &= \phi^{l,0}(\boldsymbol{h}_y^l, \boldsymbol{h}_x^l, \boldsymbol{e}_{y,x}, \hat{\boldsymbol{h}}_y, \hat{\boldsymbol{h}}_x) = \|_{k=1}^K \exp(LeakyReLU(\boldsymbol{a}^T[\hat{\boldsymbol{h}}_y[k] \parallel \hat{\boldsymbol{h}}_x[k]])), \\ \boldsymbol{s}_x^{l,0} &= \Sigma_{v_y \in \mathcal{N}(v_x)} \boldsymbol{m}_{y,x}^{l,0}, \\ \boldsymbol{h}_x^{l,0} &= \gamma^{l,0}(\boldsymbol{h}_x^l, \boldsymbol{s}_x^{l,0}) = \boldsymbol{s}_x^{l,0}. \end{aligned} \tag{1}$$

The second sub-layer of GAT (defined in Equation 2) uses the weight sum vectors to normalize the attention weights for every edge and aggregates the attention vectors $\hat{\boldsymbol{h}}_y^l$ with the normalized weights. The aggregated attention vectors $\boldsymbol{s}_x^{l,1}$ are transformed by an activation function δ and are outputted as the hidden vectors of the current layer $\boldsymbol{h}_x^{l+1}$.

$$\begin{aligned} \boldsymbol{m}_{y,x}^{l,1} &= \phi^{l,1}(\boldsymbol{h}_y^{l,0}, \boldsymbol{h}_x^{l,0}, \boldsymbol{e}_{y,x}, \hat{\boldsymbol{h}}_y, \hat{\boldsymbol{h}}_x) = \|_{k=1}^K \frac{\exp(LeakyReLU(\boldsymbol{a}^T[\hat{\boldsymbol{h}}_y[k] \parallel \hat{\boldsymbol{h}}_x[k]]))}{\boldsymbol{h}_x^{l,0}[k]} \hat{\boldsymbol{h}}_y[k], \\ \boldsymbol{s}_x^{l,1} &= \Sigma_{v_y \in \mathcal{N}(v_x)} \boldsymbol{m}_{y,x}^{l,1}, \\ \boldsymbol{h}_x^{l+1} = \boldsymbol{h}_x^{l,1} &= \gamma^{l,1}(\boldsymbol{h}^{l,0}, \boldsymbol{s}_x^{l,1}) = \delta(\boldsymbol{s}_x^{l,1}). \end{aligned} \tag{2}$$

Reviewer Point P 1.3 — Some typical applications of GNNs should be included, such as video object segmentation [ref1], human-object interaction [ref2] and human-parsing [ref3].[1] Zero-shot video object segmentation via attentive graph neural networks,iccv 2019 [2] Learning human-object interactions by graph parsing neural networks, eccv 2018. [3] Hierarchical human parsing with typed part-relation reasoning, cvpr 2020.

Reply: Thank you for pointing out our shortcomings. Computer vision is indeed another important application area of graph neural networks. We have added the mentioned references in the INTRODUCTION section in the revised manuscript. We quote the related sentence below:

Section 1

The powerful expression ability makes GNNs achieve good accuracy in not only graph analytical tasks [8, 9, 10] (like node classification and link prediction) but also computer vision tasks (like human-object interaction [11], human parsing [12], and video object segmentation [13]).

Reviewer Point P 1.4 — There are some grammar errors and typos:
- 'Take the demo GNN in Figure 1(a) as the example.'
- 'to calculate the output hidden vector h^{l+1} of the current layer l, i.e., $h^{l+1} = \gamma^l(h^l, s^l)$ The end-to-end training requires. . . '
- 'Implementing it with the specially optimized basic operators on the GPU is a potential optimization'
- The sentences in the experimental section should be unified.

Reply:

Thank you for pointing them out. We feel really sorry for our carelessness. We have proofread our revised manuscript carefully to eliminate grammar errors and typos. We have also unified the tenses of the sentences in Section 3 "Evaluation Design" and Section 4 "Evaluation Results and Analysis". We use the past tense to describe experimental methods, results and what they indicate. We only use the present tense in the sentences that Figure/Table X are the subjects of the sentences.

Reviewer Point P 1.5 — Figures 6 and 7 should be adjusted. The figures and fonts are too small.

Reply:

Thank you for pointing the problem out. Besides Figure 6 and Figure 7, we have adjusted all the figures in the revised manuscript to make sure that font sizes in figures are no less than the font size of figure captions. To make the figures more easy to read, we also splitted Figure 6 and Figure 7 into five figures (Figure 6 to 10) in the revised manuscript to enlarge the subfigures.

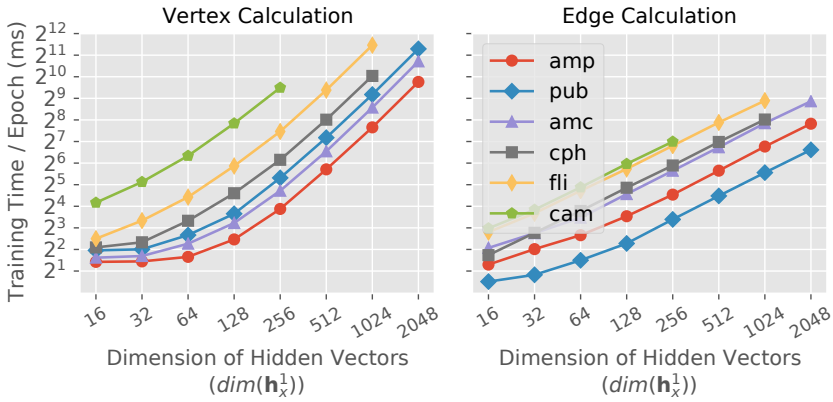
Reviewer Point P 1.6 — In my view, computation efficiency is to describe the testing or validation process. Except for reporting and analyzing the training times, it is meaningful to discuss the inference time. This is also an important point of view for deep learning researchers to be concerned about.

Reply: Thank you for the insightful comment and suggestion. The efficiency of inference (including inference time and memory usage) is indeed important for deep learning researchers and engineers. In the revised manuscript, we have added a new subsection (Section 2.5 "Inference with GNNs") in the revised manuscript to briefly review how to perform inference with GNNs. To discuss the efficiency of GNN inference on GPUs, we have added corresponding analysis on GNN inference in every subsection of Section 4 "Evaluation Results and Analysis". Below, we introduce our main findings subsection by subsection.

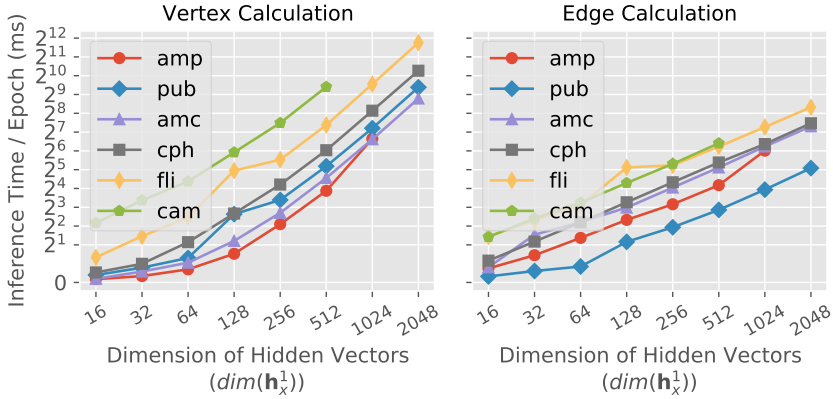
Section 4.1 "Effects of Hyper-parameters on Performance" In Section 4.1 "Effects of Hyper-parameters on Performance", we have additionally measured how the time and the peak memory usage changed as we increased the values of the hyper-parameters during *inference*.

For all GNNs, we found that the effects of hyper-parameters on the *inference time* were the same as the training time. Taking GGNN as an example, Figure 1 in the response (Figure X and Figure X in the revised manuscript) compares the effects of the hidden vector dimension $dim(\boldsymbol{h}_x^1)$ on training/inference time. As $dim(\boldsymbol{h}_x^1)$ increased, the vertex/edge calculation time of training time

and inference time both grew linearly when the values of hyper-parameters were large enough. We observed similar phenomena with the other GNNs. The results indicated that the time complexity analysis of GNNs were applicable to both training and inference.



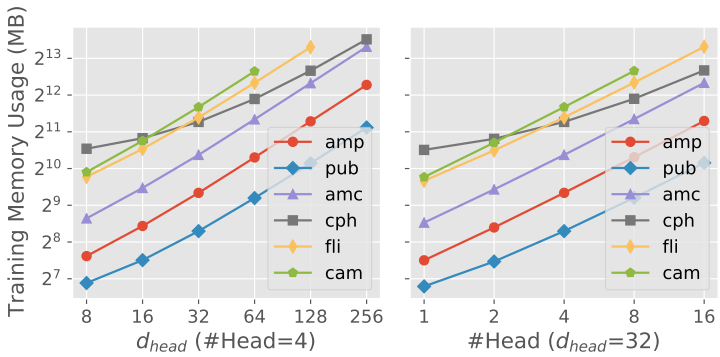
(a) Training



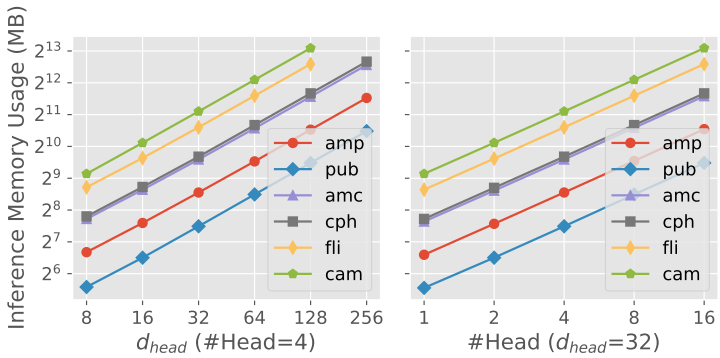
(b) Inference

Figure 1: Effects of hyper-parameters on the vertex/edge calculation time of GGNN.

For all GNNs, we found that the effects of hyper-parameters on inference memory usage were also same as training. Taking GAT as an example, Figure 2 in the response (Figure X and Figure X in the revised manuscript) compares the peak memory usage during training and inference under different values of hyper-parameters. The trends of the curves in Figure 2a and Figure 2b were highly similar. The memory usage of training and inference *both* grew linearly as the dimension of each head and the number of heads increased. We observed similar phenomena with the other GNNs.



(a) Training

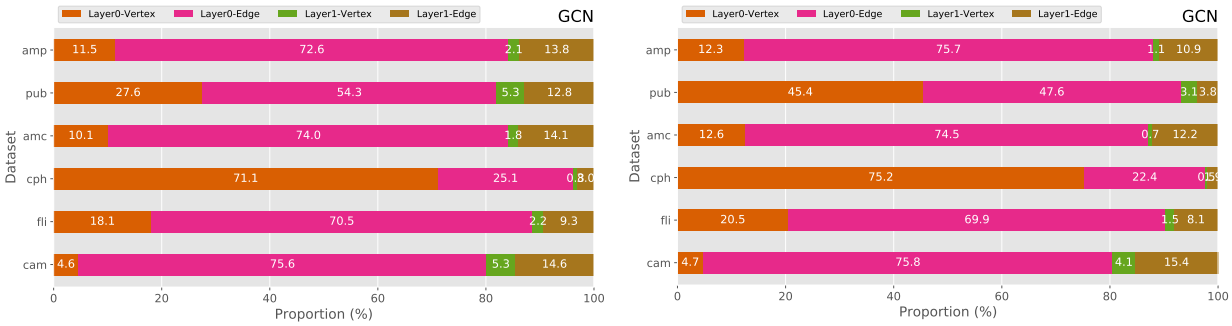


(b) Inference

Figure 2: Effects of hyper-parameters on peak memory usage of GAT.

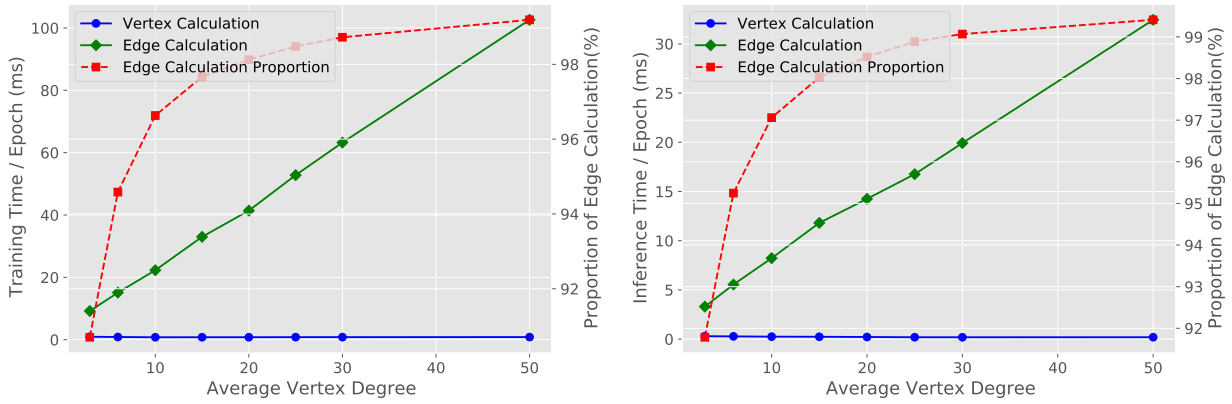
Section 4.2 "Time Breakdown Analysis" In Section 4.2 "Time Breakdown Analysis", we have additionally conducted the time breakdown analysis for GNN inference, in order to find out its performance bottlenecks.

We found that the performance characteristics of GNN inference were highly similar to GNN training on the *layer* level and the *step* level. In this reply, we used GCN as an example. We made similar observations about the other GNNs. Figure 3 in the response compares the time breakdowns of GCN on the layer level of training and inference. The time breakdowns on the layer level were very similar between inference and training. Figure 4 in the response compares the effects of the average degree on the proportion of edge/vertex calculation time during training and inference. The curves in the two sub-figures of Figure 4 show very similar trends. The results indicated that the edge calculation dominated both the training and inference time. Figure 5 in the response further compares the time breakdowns on the step level of the edge calculation stage of GCN. The proportion of each step was very similar in training and inference. Therefore, *the performance bottlenecks of training and inference were same on the layer level and step level.*



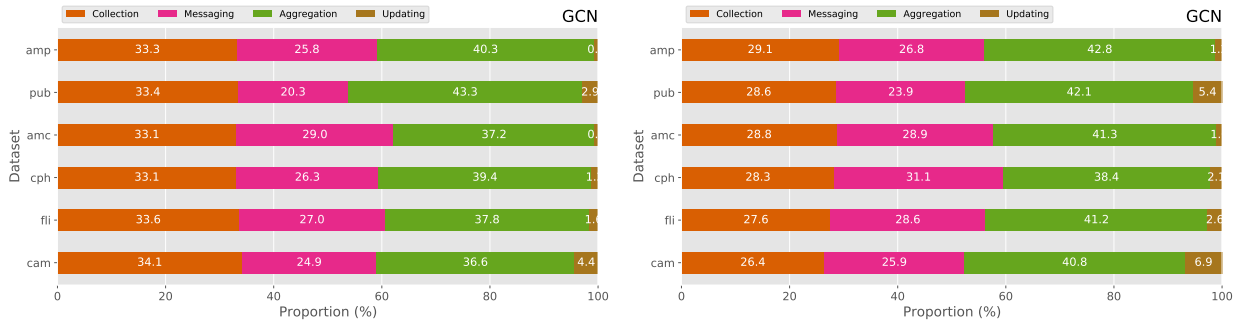
(a) Training (b) Inference

Figure 3: Time breakdowns on the layer level of GCN.



(a) Training (b) Inference

Figure 4: Effects of the average degree on the proportion of the edge/vertex calculation time of GCN. Graphs were generated by fixing the number of vertices as 50,000.

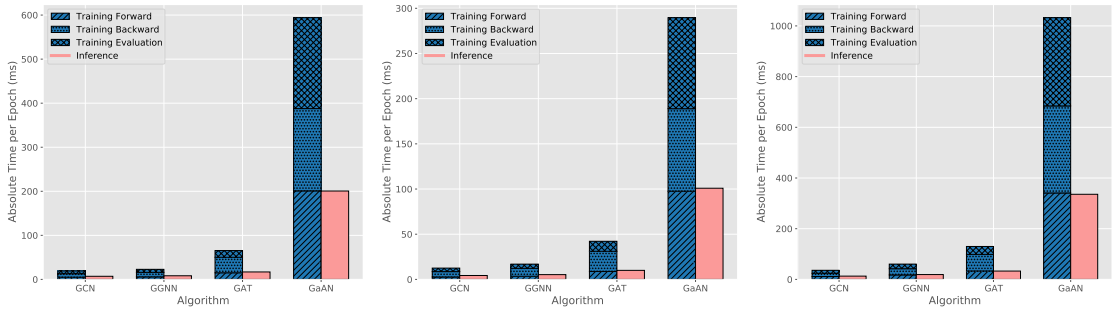


(a) Training (b) Inference

Figure 5: Time breakdowns on the step level of the edge calculation stage of GCN.

The main differences between training and inference were reflected in two aspects: the wall-clock time and the top time-consuming basic operators.

Figure 6 in the response (Figure X in the revised manuscript) compares the wall-clock training/inference time on the `amp`, `amc`, and `fli` datasets. The results on the other datasets were similar. Since the inference only conducted the forward propagation from the input layer to the prediction layer, the time of inference was very close to the time of the forward phase in training. The inference time was only 34% (GCN), 32% (GGNN), 25% (GAT), and 32% (GaAN) of the training time for the four GNNs, averaged over datasets.



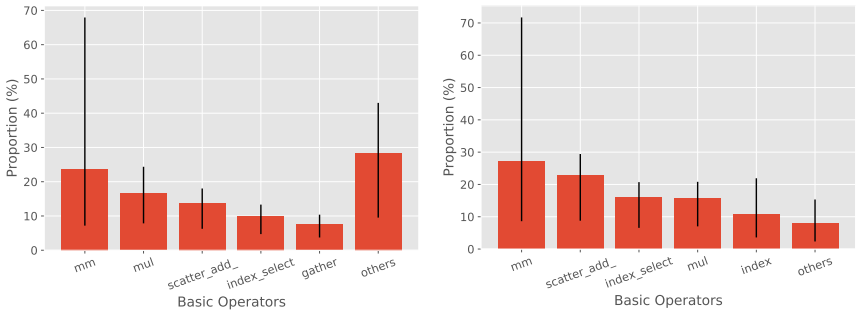
(a) `amc` (b) `amp` (c) `fli`

Figure 6: Wall-clock training/inference time on different datasets.

The top time-consuming basic operators of training and inference showed a certain degree of difference. In All GNNs, some of the top time-consuming operators during training were replaced by new operators in inference. Figure 7 in the response compares the top 5 time-consuming basic operators in training and inference. For GCN, the `index` operator that was used in the prediction layer became the new top 5 time-consuming operator, replacing the `gather` operator used in the backward phase in training. The basic operators related to the edge calculation (`scatter_add`, `index_select`, and `mul`) still consumed the majority of the inference time. For GGNN, `index`

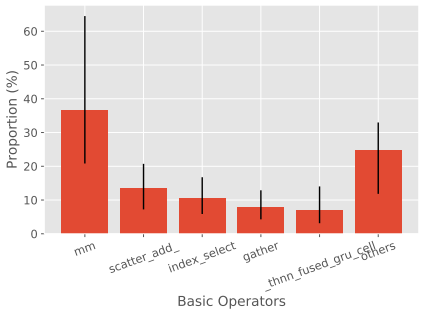
operator in the prediction layer also became one of the top 5 time-consuming basic operators. Due to the high time complexity of the vertex updating function γ in GGNN, the basic operators related to the vertex calculation stage (`mm` and `thnn_fused_gru_cell`) still consumed near half of the inference time. For GAT, the `input_put_impl` operator (used in the backward phase) in Figure 7e was replaced by the `scatter_add` operator in Figure 7f. The `scatter_add` operator was still related to the edge calculation step. For GaAN, the `mm` operator of GaAN was replaced by the `cat` operator used in the vertex updating function.

Comparing the top basic operator distribution, GNN training and inference had common performance bottlenecks on the operator level. The matrix multiplication `mm` and the element-wise multiplication `mul` operators were the common time-consuming operators, making GNN training and inference *both suitable for GPUs*. The basic operators related to the collection and aggregation steps in the edge calculation consumed consumed non-trivial time in both training and inference.

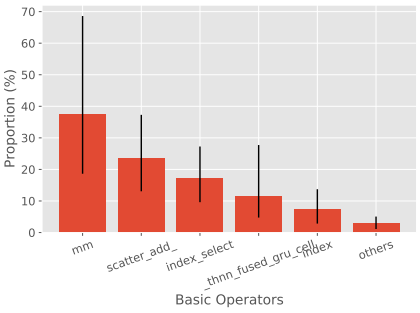


(a) GCN, Training

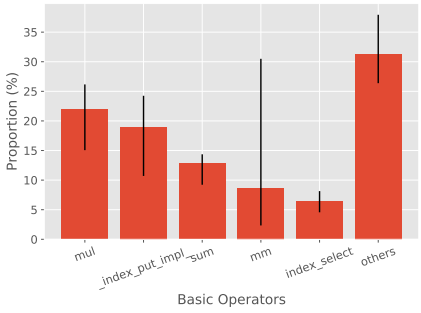
(b) GCN, Inference



(c) GGNN, Training

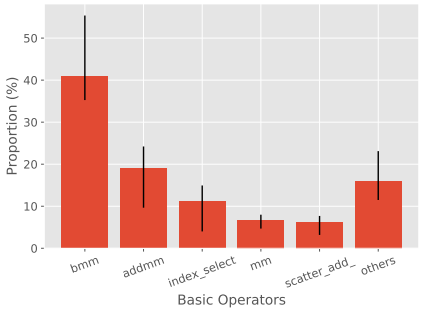


(d) GGNN, Inference

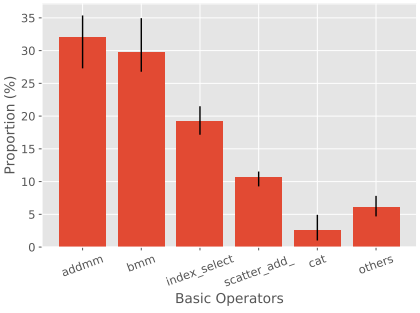


(e) GAT, Training

(f) GAT, Inference



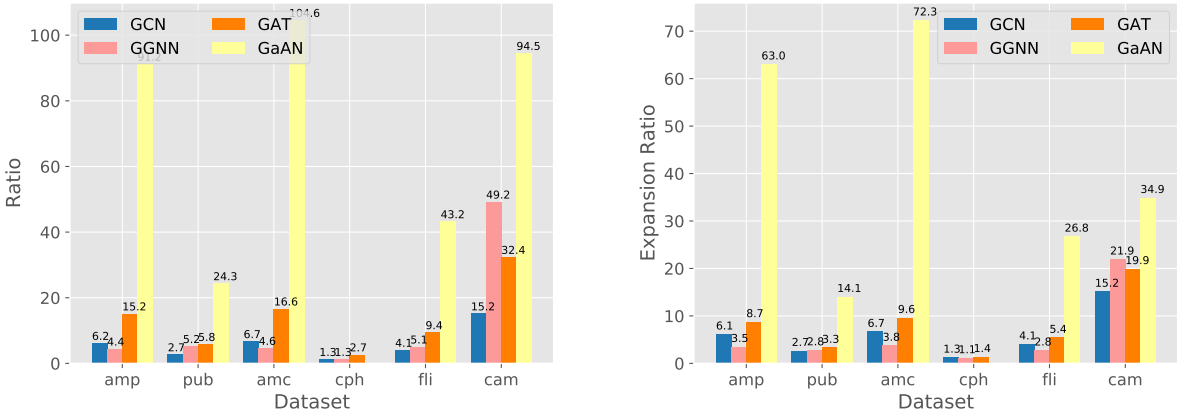
(g) GaAN, Training



(h) GaAN, Inference

Figure 7: Top 5 time-consuming basic operators of typical GNNs. The time proportion of each basic operator was averaged over all datasets with the error bar indicating the maximum and the minimum.

Section 4.3 "Memory Usage Analysis" In Section 4.3 "Memory Usage Analysis", we have additionally conducted the memory usage analysis for GNN inference and added related discussion. Figure 8 in the response (Figure X in the revised manuscript) compares the memory expansion ratios of typical GNNs during training and inference. Since no intermediate results had to be cached during inference, the memory expansion ratios of inference were much less than training for GGNN, GAT, and GaAN. The MERs of inference were 45% to 83% (GGNN), 52% to 61% (GAT), and 37% to 69% (GaAN) of training. However, the values of MERs were still high, disallowing inferencing with big graphs.



(a) Training (b) Inference

Figure 8: Memory expansion ratios of typical GNNs.

Section 4.4 "Effects of Sampling Techniques on Performance" In Section 4.4 "Effects of Sampling Techniques on Performance", we have additionally evaluated the performance bottlenecks in sample-based GNN inference. We find that ...

Reviewer 2

Reviewer Point P 2.1 — In the paper, authors accomplished a unique study and analysis on GNN models training complexity. The articles first review and development history of GNNs and creatively model all architectures as input layers, intermediate layers of graph neurons and prediction layers. And they quantitatively summarize the time and space complexity of 4 representative GNNs, including graph convolution, gated recurrent graph net, graph attention net and GraphSage. Most importantly, the article first break down complexity into operator level and offered analysis of good granularity, giving reader more guidance in future study. At last, the solid experiments included the study of effects of

hyper-parameters and a comparison of two major sampling techniques: neighbor sampling and cluster sampling.

Reply: Thank you for your positive comments on our manuscript. We have carefully revised the manuscript according to your comments. We have revised our manuscript according to your kindly suggestions. Please see the detailed responses below. We have highlighted our modification point by point in the annotated version of the manuscript by red squares.

Reviewer Point P 2.2 — In general, the paper was well written and organized with good structure and clear narratives. Just some minor language errors like line Page 8, Line 208, "In active graph neurons" =>"Inactive graph neurons".

Reply: Thank you for pointing them out. We feel really sorry for our carelessness. We have proofread our revised manuscript carefully to eliminate such language errors as much as we can.

Reviewer Point P 2.3 — I was impressed by the way that authors categorize layers and operators in GNNs, very clear and instructive.

It is also pretty neat to divide layer time complexity into two buckets: vertex calculation and edge calculation. The data model pretty well summarizes mainstream GNN layer architectures. And this analysis is very insightful for layer profiling.

And the experimental evaluation were done over 6 large graph-structured datasets.

Reply: Thank you very much for your appreciation.

Reviewer Point P 2.4 — While, one major drawback is that I did not clearly see the analysis complexity v.s. accuracy. For example, in Figure 19 and 20, I did not see network accuracy from those 4 GNNs. There is always tradeoff between model complexity and model performance, and in some scenarios where high complexity is allowed, a sophisticated model of more powerful representation capability is still needed.

Reply: Thanks very much for your valuable suggestions. The model complexity directly affected both the accuracy and the training time. In order to analyze the relationship between model complexity and accuracy, we conducted two kinds of extra experiments in the revised manuscript: (1) how the hyper-parameters of the GNNs (like the dimension of hidden vectors and the number of heads) affected the accuracy of GNNs (in Section 4.1); (2) how the batch size in the sampling methods affected the accuracy of GNNs (in Section 4.4).

In this reply, we focus on the first kind of experiments added in Section 4.1. We will introduce our results of the second kind of experiments in the nextR reply.

In the revised manuscript, we added an extra paragraph "Effects on Accuracy" at the end of Section 4.1 "Effects of Hyper-parameters on Performance" to analyze how the hyper-parameters affect the accuracy. We had two main findings: (1) the accuracy of GNNs was much more sensitive to the dimension of hidden vectors $dim(\boldsymbol{h}_x^1)$ (for GCN/GGNN/GaAN) and the dimension of each head d_{head} (for GAT) than the other hyper-parameters; (2) the relative accuracy between the four typical GNNs varied greatly with different datasets. We quote the related paragraphs from the revised manuscript below:

To add the source from manuscript

Reviewer Point P 2.5 — Sampling method is definitely going to reduce model complexity, since all models complexity depend on graph node number N, while performance is going to be compromised as well. I would like to see authors resolve the concern of significant accuracy drop after applying aggressive sampling of subgraphs.

Reply:

Reviewer Point P 2.6 — Hope authors supplement the effect of sampling and GNNs on accuracy while comparing different complexity of model and sampling methods.

Reply: