Response to the Reviewers

We thank the reviewers for their insightful comments on our work. We have made modifications to our manuscript according to Reviewer 1's comments.

Reviewer 1

Reviewer Point P 1.1 — In the revised manuscript, the authors have addressed all my concerns, especially adding more computation efficiency discussion. Current version looks very impressive. Just a small issue left. Some discussions about limitations or future efforts should be better added.

Reply: Thank you for the insightful suggesion. According to the suggestion, we have added a paragraph at the end of Section 7 "Conclusion and Future Work" to discuss limitations of our work and points out potential future research directions. In this work, we mainly analyze performance bottlenecks of GNN training/inference in a *single-GPU* environment on *static* graphs with the *message-passing* framework. Performance bottlenecks in *multi-GPU/distributed* GNN training/inference with *dynamic* graphs and *other* GNN frameworks are also worth studying. In the future, we plan to extend our work in the following directions:

1. *Multi-GPU and distributed GNN training/inference.* To handle large-scale graph datasets, training/inferring GNNs with multiple GPUs or in a distributed environment is necessary. Multi-GPU and distributed GNN training/inference will inevitably introduce overheads such as inter-GPU and inter-machine communication. How these overheads affect performance bottlenecks is worthy to focus on.

2. *Spatial-temporal graph datasets.* Spatial-temporal graphs have dynamic topology structures. They appear in a variety of applications like traffic speed forecasting [Li et al. (2018)] and human action recognition [Yan et al. (2018)]. Many new GNNs are proposed to handle this kind of dynamic graphs. How the performance bottlenecks of these GNNs are different from the classic GNNs is also worthy of in-depth study.

3. *Other GNN frameworks.* In this work, we conducted analysis with the message-passing framework that is popular among existing GNN learning systems. Some emerging GNN learning systems also adopt different frameworks like SAGA framework [Ma et al. (2019)] and edge-centric framework [He (2019)]. Whether different frameworks lead to different performance bottlenecks is worth further investigation.

References

He, L. (2019). Engn: A high-throughput and energy-efficient accelerator for large graph neural networks. *CoRR*, abs/1909.00155.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Ma, L., Yang, Z., Miao, Y., Xue, J., Wu, M., Zhou, L., and Dai, Y. (2019). Neugraph: Parallel deep neural network computation on large graphs. pages 443–458. 2019 USENIX Annual Technical Conference, USENIX ATC 2019, Renton, WA, USA, July 10-12, 2019.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7444–7452. AAAI Press.