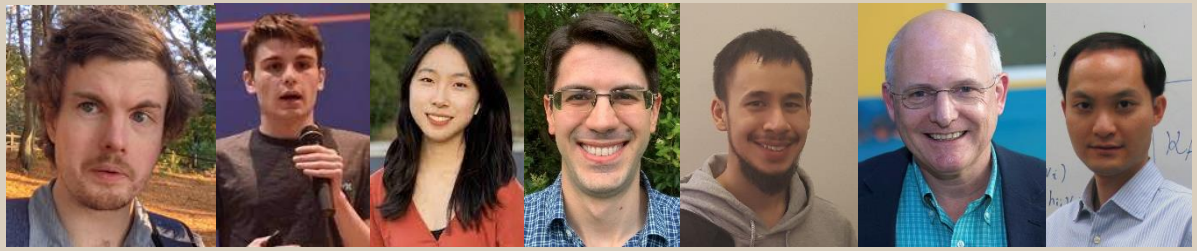
Fast and Scalable Graph Neural Networks with Pytorch + SALIENT

based on

Accelerating Training and Inference of Graph Neural Networks with Fast Sampling and Pipelining – MLSys 2022



Tim Kaler*, Nick Stathas*, Anne Ouyang*, Alexandros-Stavros Iliopoulos, Tao B. Schardl, Charles E. Leiserson, Jie Chen

May 17th 2022

* Equal contribution

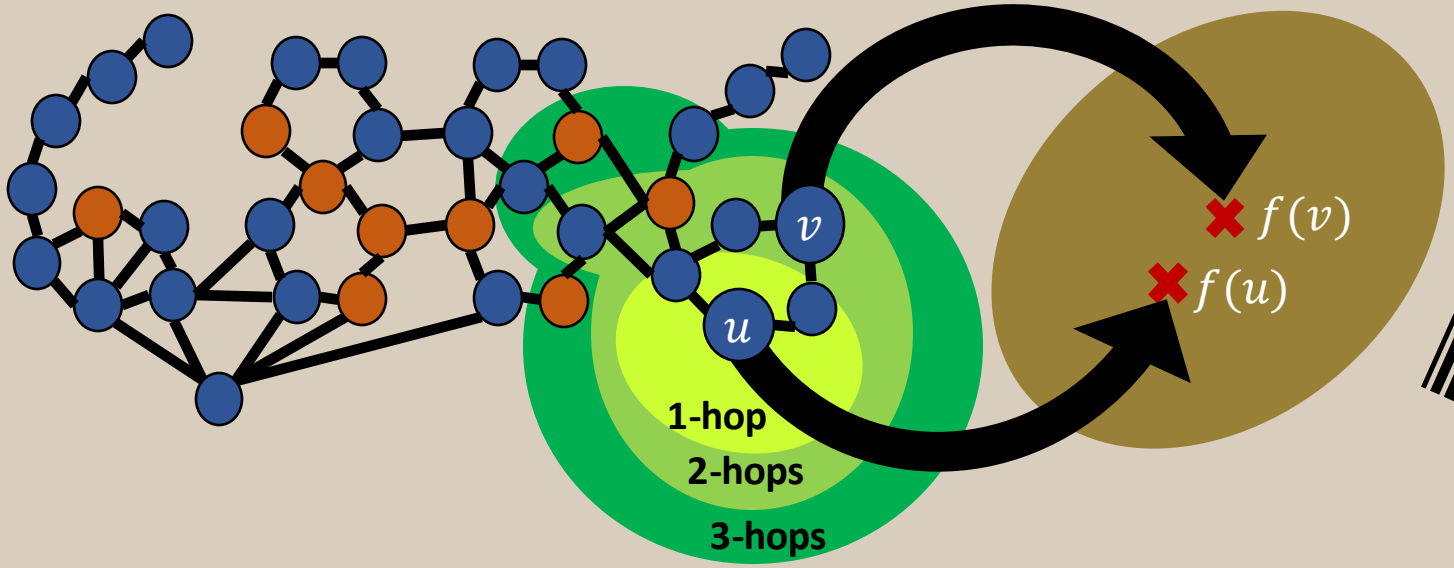
MIT-IBM Watson AI Lab

CSAIL

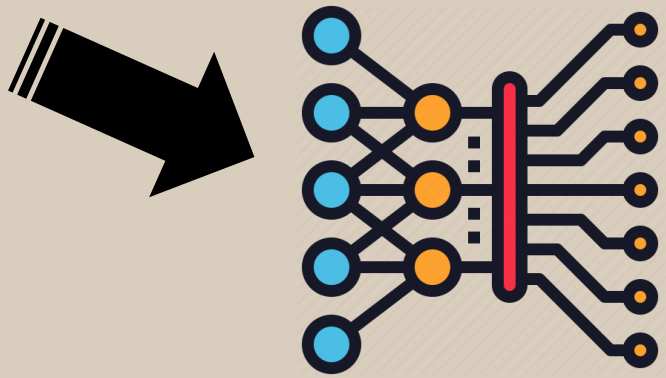
GNNs learn embeddings that account for graph structure

Embed nodes into a lower dimensional space.

$$f : V \rightarrow R^N$$



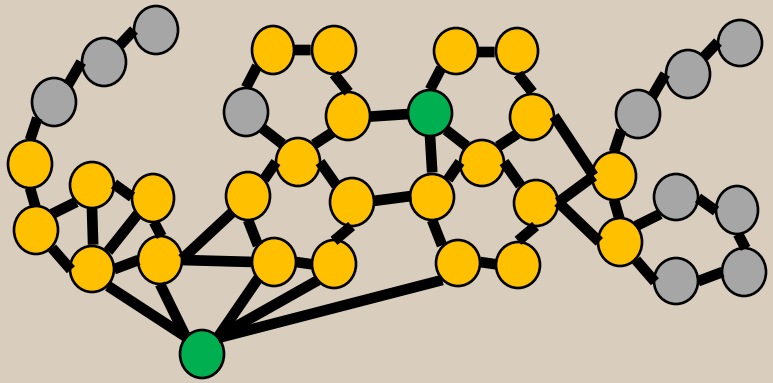
Embeddings fed into network



GNNs can learn an embedding function that encodes the **k**-hop neighborhood of a vertex.

Neighborhood Sampling in Graph Neural Networks

Mini-batch training on GNNs: Mini-batch training and stochastic gradient descent substantially improves training performance.



- Mini-batch node
- 3-hop neighborhood
- Outside of 3-hop neighborhood

Problem

The expanded neighborhood of even a small mini-batch may be a sizeable portion of whole graph!

Solution

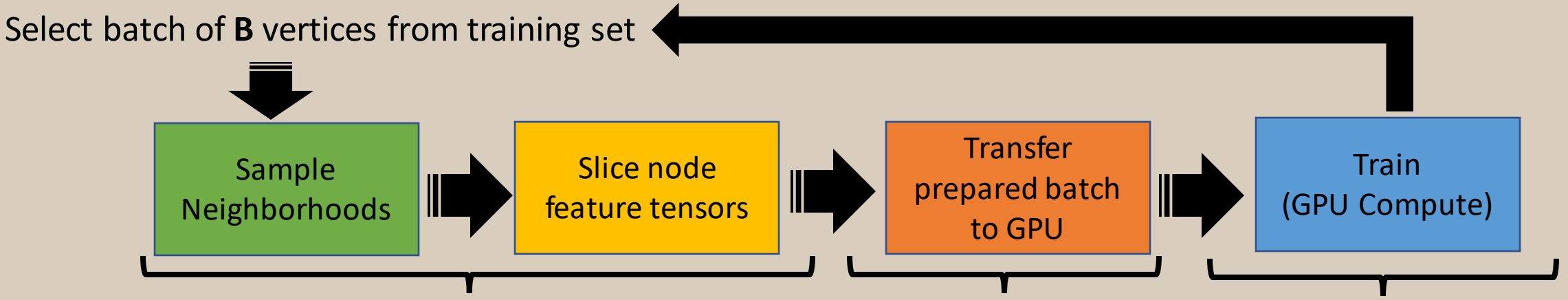
Neighborhood sampling is used to reduce expanded neighborhood size.

But…

New problem

Neighborhood sampling often bottlenecks GNN training

Performance breakdown of GNN training in PyTorch



	Total	Batch Preparation		Data Transfer		Train (GPU)	
Benchmark		*Time*	*%*	*Time*	*%*	*Time*	*%*
arxiv	1.7 sec	1.0 sec	58%	0.3 sec	15%	0.5 sec	27%
products	8.6 sec	4.0 sec	46%	2.2 sec	26%	2.4 sec	28%
papers	50.4 sec	18.6 sec	37%	17.9 sec	35%	13.9 sec	28%

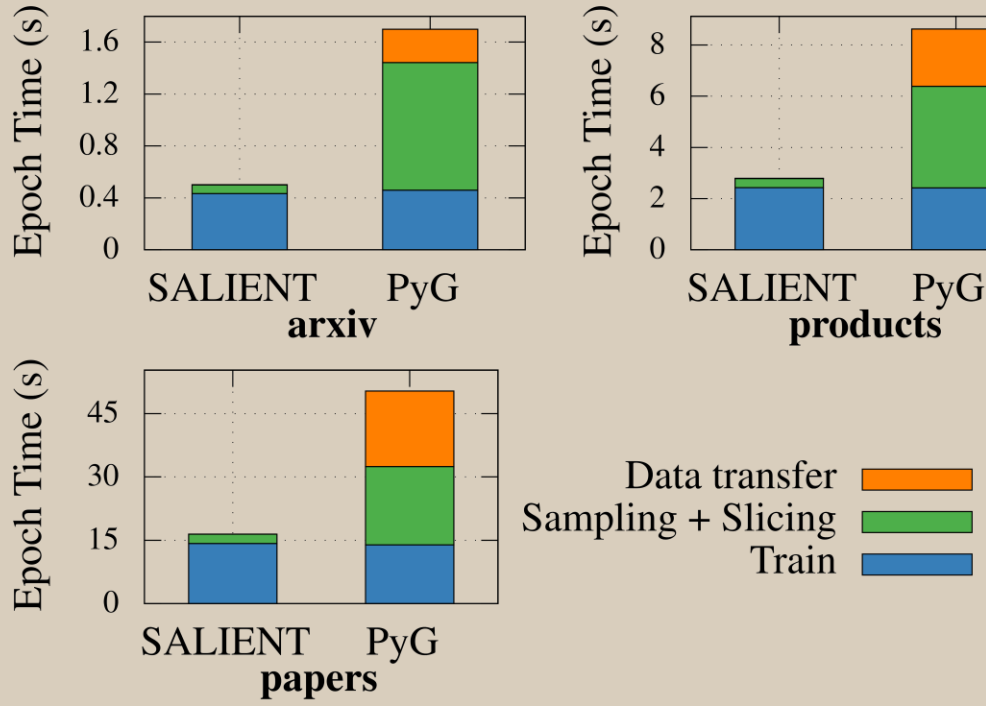
Less than 30% of per-epoch time is spent waiting on GPU compute

Improving GNN training performance with SALIENT

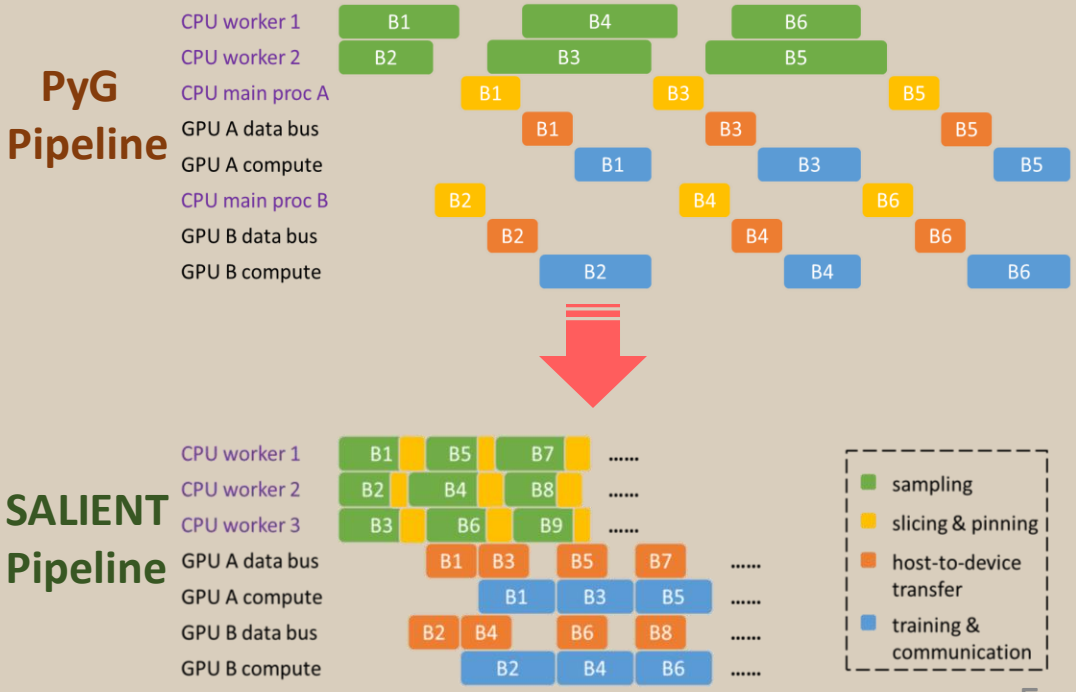
SALIENT accelerates GNN training with **fast sampling + pipelining + shared-memory parallelism**

Performance benefits of SALIENT do not require adopting a new ML framework or rewriting GNN models. SALIENT can be used as a drop-in replacement for the NeighborSampler in Pytorch Geometric.

Performance breakdown of SALIENT and PyG on 1 GPU

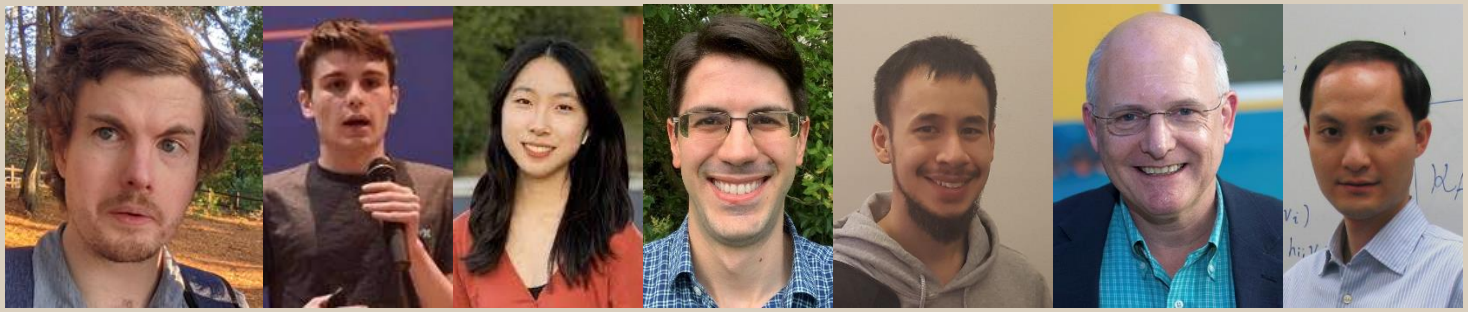


Comparison of PyG and SALIENT pipelines



Visit breakout room or see full paper for more details of our performance study and the design of SALIENT

SALIENT and xGraph Systems Team



Tim Nick Anne Alex TB Charles Jie

Obada Phil

Full Paper

"Accelerating Training and Inference of Graph Neural Networks with Fast Sampling and Pipelining" to appear in MLSys 2022 (https://arxiv.org/abs/2110.08450)

Software

https://github.com/MITIBMxGraph/SALIENT_artifact

https://github.com/MITIBMxGraph/SALIENT

Acknowledgements

This research was sponsored by MIT-IBM Watson AI Lab and in part by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.