Graph-less Neural Networks:

Teaching Old MLPs New Tricks via Distillation

Shichang Zhang[1], Yozen Liu[2], Yizhou Sun[1], Neil Shah[2]

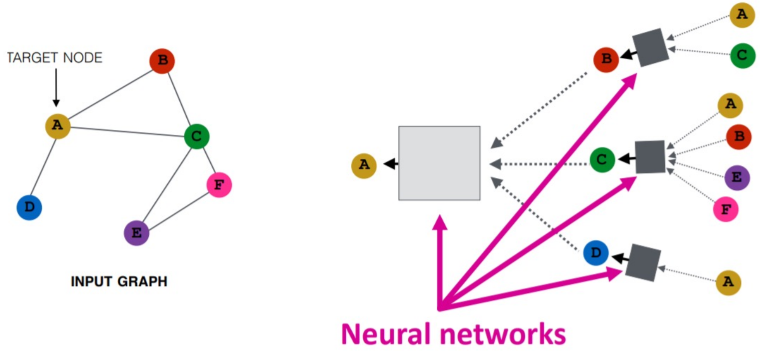
[1]University of California, Los Angeles (UCLA) [2]Snap Inc.

May 2022

UCLA

Snap Inc.

Graph Neural Network (GNN) in Production



GNNs

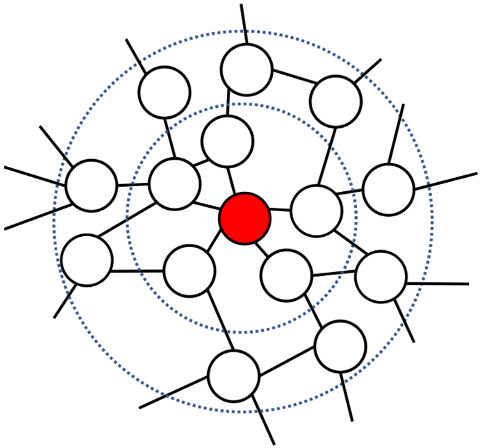
- Message passing between neighbor nodes, which is a recursive process extends to multi-hop neighbors

Industrial applications

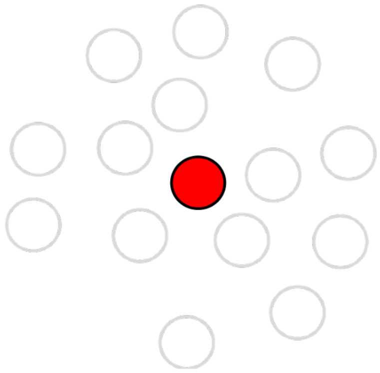
- Large scale graph data
- Expensive neighbor fetching
- Latency-constrained tasks
- Multi-layer Perceptron (MLP) remains the major workhorse

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GNN vs. Multi-layer Perceptron (MLP)



GNN: message passing between data points



MLP: independence between data points

GNN vs. MLP: Accuracy

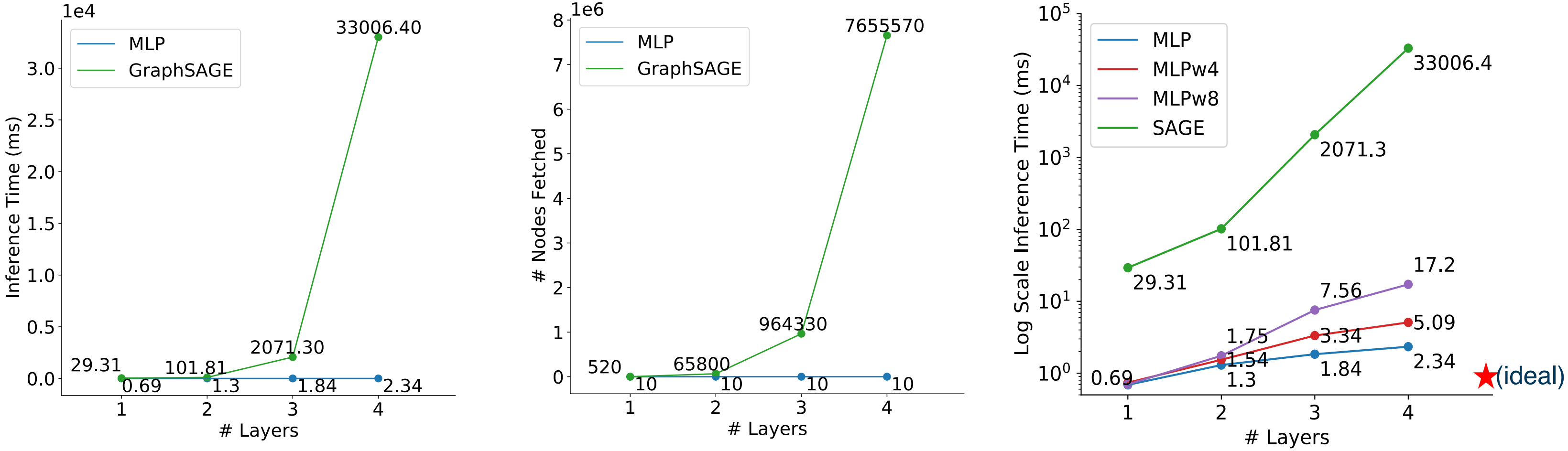
Datasets	GraphSAGE	MLP
Cora	80.52 ± 1.77	59.22 ± 1.31
Citeseer	70.33 ± 1.97	59.61 ± 2.88
Pubmed	75.39 ± 2.09	67.55 ± 2.31
A-computer	82.97 ± 2.16	67.80 ± 1.06
A-photo	90.90 ± 0.84	78.77 ± 1.74
Arxiv	70.92 ± 0.17	56.05 ± 0.46
Products	78.61 ± 0.49	62.47 ± 0.10

Node classification accuracy on seven benchmarks

Accuracy of GNN (GraphSAGE) significantly outperforms MLP

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GNN vs. MLP: Inference Time

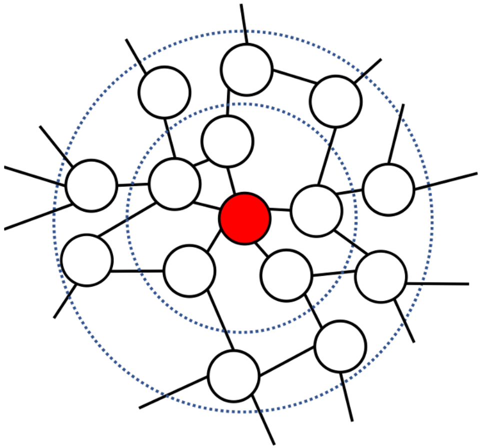


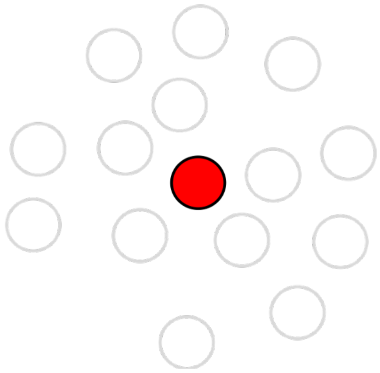
Infer 10 randomly selected nodes (Products graph, ~2.5M nodes)
time = **fetching data** + forward pass

- **GNN**: Node fetching causes inference time to grow exponentially with respect to # layers
- **MLP**: Inference time grows only linearly and remains much smaller than GNNs even with more parameters.

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GNN vs. Multi-layer Perceptron (MLP)





GNN: message passing between data points

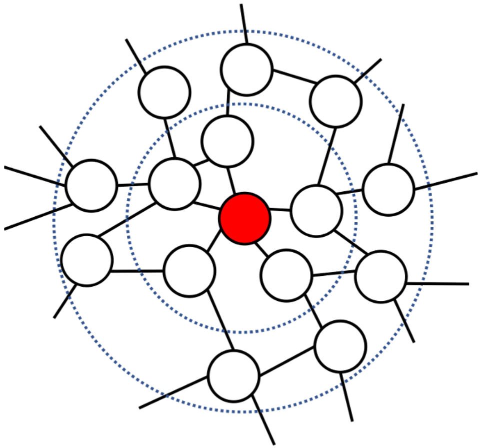
- High accuracy
- Graph dependency (neighbor fetching)
  - Deployment challenge
  - Inference latency

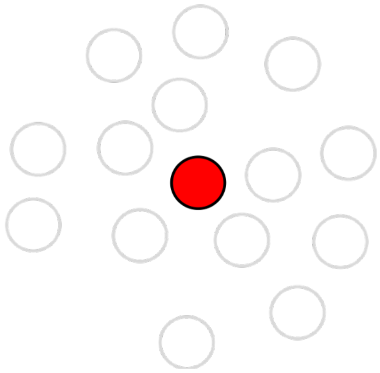
MLP: independence between data points

- Less accurate than GNN
- No graph dependency
  - Faster and easier to deploy
  - Sidestep the cold-start problem

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GNN and MLP: Combine Advantages





Accurate GNN:

- Graph dependency in learning
- Graph dependency in inference

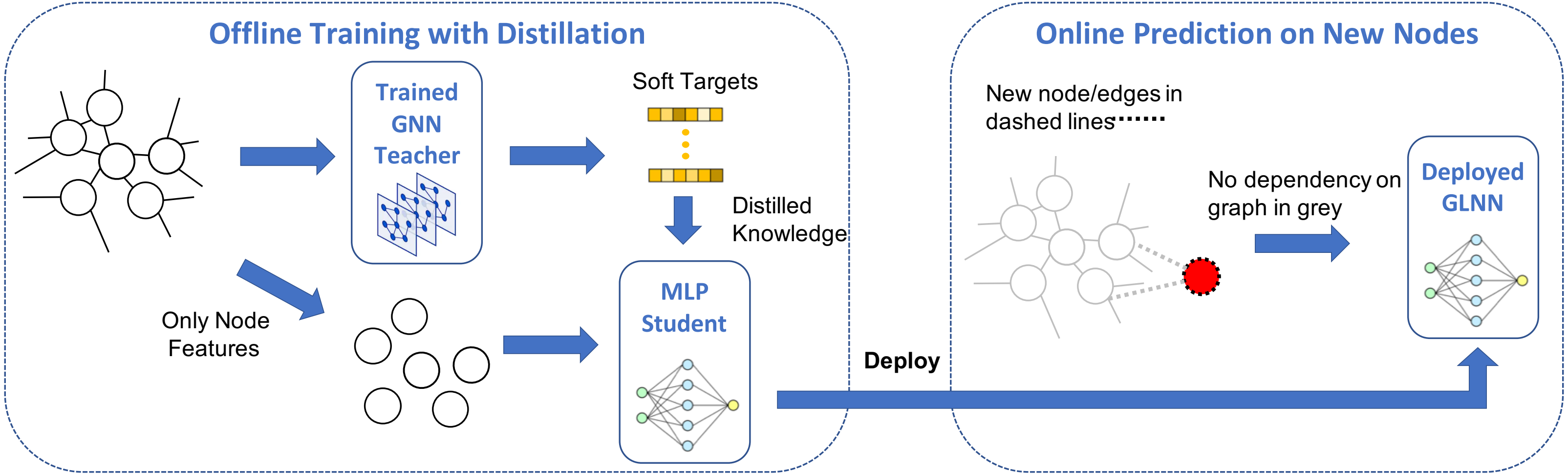
Fast MLP:

- No graph dependency in learning
- No graph dependency in inference

Can we use graph dependency in learning, but not inference?

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

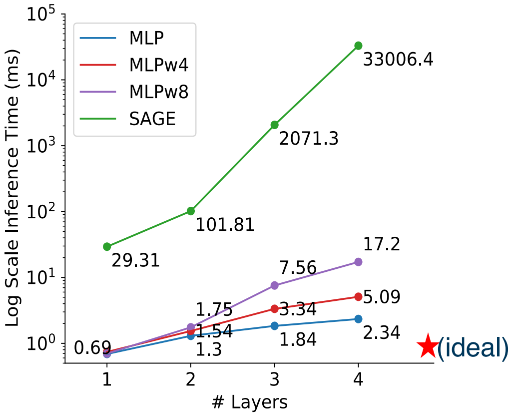
Our Proposal: Graph-less Neural Network (GLNN)



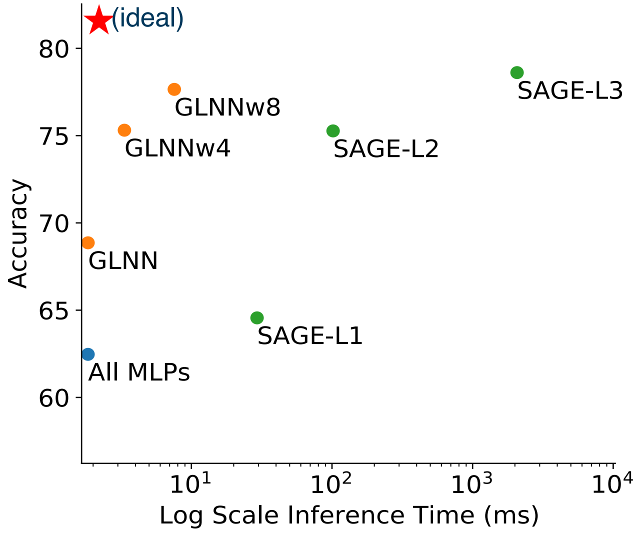
- **Offline training:** graph-dependent GNN + knowledge distillation (KD) to MLP
- **Online prediction:** faster and more accurate inference for new nodes

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

Trade-offs Between Speed and Accuracy



Add accuracy



- GLNN accuracy improves **greatly** from MLP
- GLNNs are **much faster** and **comparably accurate** to GNN

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GLNN Results: Accuracy

Datasets	Eval	SAGE	MLP/MLP+	GLNN/GLNN+	Δ_{MLP}	Δ_{GNN}
Cora	*prod*	**79.29**	58.98	78.28	19.30 (32.72%)	-1.01 (-1.28%)
	ind	81.33 ± 2.19	59.09 ± 2.96	73.82 ± 1.93	14.73 (24.93%)	-7.51 (-9.23%)
	tran	78.78 ± 1.92	58.95 ± 1.66	79.39 ± 1.64	20.44 (34.66%)	0.61 (0.77%)
Citeseer	*prod*	68.38	59.81	**69.27**	9.46 (15.82%)	0.89 (1.30%)
	ind	69.75 ± 3.59	60.06 ± 5.00	69.25 ± 2.25	9.19 (15.30%)	-0.5 (-0.7%)
	tran	68.04 ± 3.34	59.75 ± 2.48	69.28 ± 3.12	9.63 (15.93%)	1.24 (1.82%)
Pubmed	*prod*	**74.88**	66.80	74.71	7.91 (11.83%)	-0.17 (-0.22%)
	ind	75.26 ± 2.57	66.85 ± 2.96	74.30 ± 2.61	7.45 (11.83%)	-0.96 (-1.27%)
	tran	74.78 ± 2.22	66.79 ± 2.90	74.81 ± 2.39	8.02 (12.01%)	0.03 (0.04%)
A-computer	*prod*	82.14	67.38	**82.29**	14.90 (22.12%)	0.15 (0.19%)
	ind	82.08 ± 1.79	67.84 ± 1.78	80.92 ± 1.36	13.08 (19.28%)	-1.16 (-1.41%)
	tran	82.15 ± 1.55	67.27 ± 1.36	82.63 ± 1.40	15.36 (22.79%)	0.48 (0.58%)
A-photo	*prod*	91.08	79.25	**92.38**	13.13 (16.57%)	1.30 (1.42%)
	ind	91.50 ± 0.79	79.44 ± 1.72	91.18 ± 0.81	11.74 (14.78%)	-0.32 (-0.35%)
	tran	90.80 ± 0.77	79.20 ± 1.64	92.68 ± 0.56	13.48 (17.01%)	1.70 (1.87%)
Arxiv	*prod*	**70.73**	55.30	65.09	9.79 (17.70%)	-5.64 (-7.97%)
	ind	70.64 ± 0.67	55.40 ± 0.56	60.48 ± 0.46	4.3 (7.76%)	-10.94 (-15.49%)
	tran	70.75 ± 0.27	55.28 ± 0.49	71.46 ± 0.33	11.16 (20.18%)	-4.31 (-6.09%)
Products	*prod*	**76.60**	63.72	75.77	12.05 (18.91%)	-0.83 (-1.09%)
	ind	76.89 ± 0.53	63.70 ± 0.66	75.16 ± 0.34	11.44 (17.96%)	-1.73 (-2.25%)
	tran	76.53 ±0.55	63.73 ± 0.69	75.92 ± 0.61	12.20 (19.15%)	-0.61 (-0.79%)

Significant accuracy improvement over MLPs.

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GLNN Results: Accuracy

Datasets	Eval	SAGE	MLP/MLP+	GLNN/GLNN+	Δ_{MLP}	Δ_{GNN}
`Cora`	*prod*	**79.29**	58.98	78.28	19.30 (32.72%)	-1.01 (-1.28%)
	ind	81.33 ± 2.19	59.09 ± 2.96	73.82 ± 1.93	14.73 (24.93%)	-7.51 (-9.23%)
	tran	78.78 ± 1.92	58.95 ± 1.66	79.39 ± 1.64	20.44 (34.66%)	0.61 (0.77%)
`Citeseer`	*prod*	68.38	59.81	**69.27**	9.46 (15.82%)	0.89 (1.30%)
	ind	69.75 ± 3.59	60.06 ± 5.00	69.25 ± 2.25	9.19 (15.30%)	-0.5 (-0.7%)
	tran	68.04 ± 3.34	59.75 ± 2.48	69.28 ± 3.12	9.63 (15.93%)	1.24 (1.82%)
`Pubmed`	*prod*	**74.88**	66.80	74.71	7.91 (11.83%)	-0.17 (-0.22%)
	ind	75.26 ± 2.57	66.85 ± 2.96	74.30 ± 2.61	7.45 (11.83%)	-0.96 (-1.27%)
	tran	74.78 ± 2.22	66.79 ± 2.90	74.81 ± 2.39	8.02 (12.01%)	0.03 (0.04%)
`A-computer`	*prod*	82.14	67.38	**82.29**	14.90 (22.12%)	0.15 (0.19%)
	ind	82.08 ± 1.79	67.84 ± 1.78	80.92 ± 1.36	13.08 (19.28%)	-1.16 (-1.41%)
	tran	82.15 ± 1.55	67.27 ± 1.36	82.63 ± 1.40	15.36 (22.79%)	0.48 (0.58%)
`A-photo`	*prod*	91.08	79.25	**92.38**	13.13 (16.57%)	1.30 (1.42%)
	ind	91.50 ± 0.79	79.44 ± 1.72	91.18 ± 0.81	11.74 (14.78%)	-0.32 (-0.35%)
	tran	90.80 ± 0.77	79.20 ± 1.64	92.68 ± 0.56	13.48 (17.01%)	1.70 (1.87%)
`Arxiv`	*prod*	**70.73**	55.30	65.09	9.79 (17.70%)	-5.64 (-7.97%)
	ind	70.64 ± 0.67	55.40 ± 0.56	60.48 ± 0.46	4.3 (7.76%)	-10.94 (-15.49%)
	tran	70.75 ± 0.27	55.28 ± 0.49	71.46 ± 0.33	11.16 (20.18%)	-4.31 (-6.09%)
`Products`	*prod*	**76.60**	63.72	75.77	12.05 (18.91%)	-0.83 (-1.09%)
	ind	76.89 ± 0.53	63.70 ± 0.66	75.16 ± 0.34	11.44 (17.96%)	-1.73 (-2.25%)
	tran	76.53 ±0.55	63.73 ± 0.69	75.92 ± 0.61	12.20 (19.15%)	-0.61 (-0.79%)

Competitive accuracy to GNNs on 6/7 datasets.

GLNN+:
GLNNw4 on ArXiv:
~160,000 nodes and ~1.1M edges
GLNNw8 on Products:
~2.5M nodes and ~61M edges

GLNN Results: Inference Time

Compare GLNN inference time to other common inference acceleration methods

- **SAGE:** Base GNN model
- **QSAGE**: Quantized SAGE, FP32 to INT8
- **PSAGE:** Pruned SAGE, with 50% model parameters pruned
- **Neighbor Sampling:** sampling 15 nodes per layer

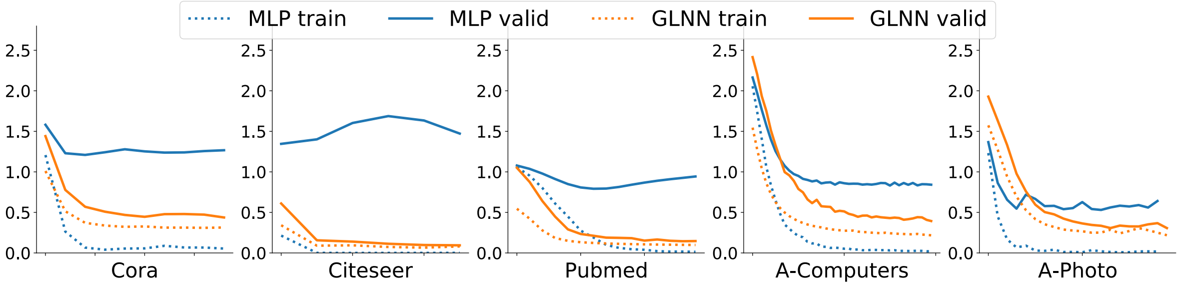
Table 4: Common inference acceleration methods speed up SAGE, but still considerably slower than GLNNs. Numbers (in *ms*) are inductive inference on 10 randomly chosen nodes.

Datasets	SAGE	QSAGE	PSAGE	Neighbor Sample	GLNN+
Arxiv	489.49	433.90 (1.13×)	465.43 (1.05×)	91.03 (5.37×)	**3.34 (146.55×)**
Products	2071.30	1946.49 (1.06×)	2001.46 (1.04×)	107.71 (19.23×)	**7.56 (273.98×)**

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

How Does GLNN Benefit from KD?

KD helps to **regularize** training of the MLP and mitigates overfitting.



The classification loss on **true labels**. GLNN curves exclude KD loss.

KD helps MLPs to **match inductive bias** of GNNs.

$$\mathcal{L}_{cut} = \frac{Tr(\hat{\mathbf{Y}}^T \boldsymbol{A} \hat{\mathbf{Y}})}{Tr(\hat{\mathbf{Y}}^T \boldsymbol{D} \hat{\mathbf{Y}})} \qquad \mathcal{L}_{cut} \in [0,1]$$

measures consistency between model prediction ($\hat{\mathbf{Y}}$) and graph topology ($\boldsymbol{A}$: adjacency matrix, $\boldsymbol{D}$: degree matrix)

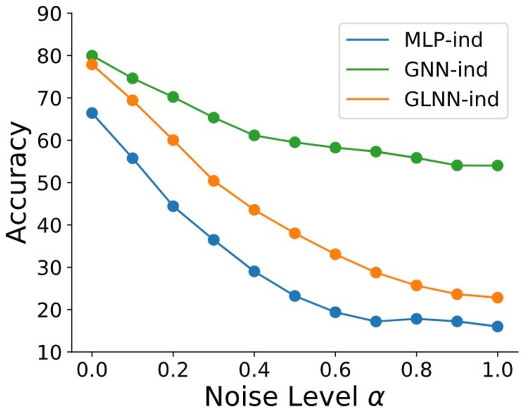
Datasets	SAGE	MLP	GLNN
Cora	0.9347	0.7026	0.8852
Citeseer	0.9485	0.7693	0.9339
Pubmed	0.9605	0.9455	0.9701
A-computer	0.9003	0.6976	0.8638
A-photo	0.8664	0.7069	0.8398
Average	0.9221	0.7644	0.8986

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

When Does GLNN Fail?

GLNNs are **less useful** in cases where labels have low correlation with node features. For example, they may be more related to the structure roles, like using node degrees as labels

Add Gaussian noise to node features $\tilde{\boldsymbol{X}} = (1-\alpha)\boldsymbol{X} + \alpha\epsilon$



- As the correlation between labels and node features decreases
  - GNN maintains reasonable prediction accuracy utilizing graph structure information
  - GLNN gets less accurate but still better than standalone MLP

NB: In practical tasks, the node features and structural roles are often highly correlated (Lerique et al. 2020).

Future Work

- Students with limited node fetching
- More sophisticated distillation techniques
- A guiding principle to decide whether GLNN is applicable to a given graph
- Towards the cold start problem as in Zheng et al. (2022)

Thank you!
Q & A

Paper link

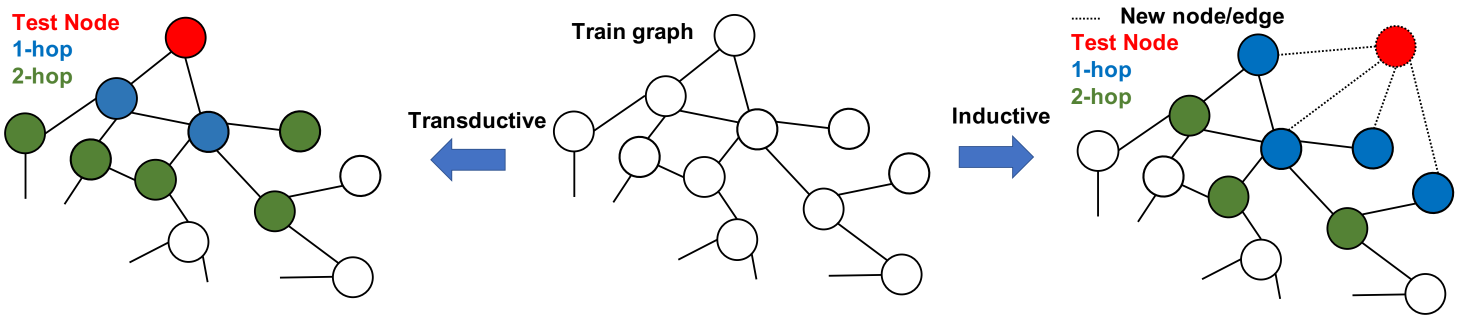
Contact author

Reference

- Lerique, S., Abitbol, J. L., & Karsai, M. (2020). Joint embedding of structure and features via graph convolutional networks. *Applied Network Science*, *5*(1), 1-24.
- Zheng, W., Huang, E. W., Rao, N., Katariya, S., Wang, Z., & Subbian, K. (2021). Cold Brew: Distilling Graph Node Representations with Incomplete or Missing Neighborhoods.
- GNN illustration picture: https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-neural-networks

Appendix

Transductive vs. Inductive



Test nodes in the transductive setting: node features and structures have been observed during training, but labels are not.

Test nodes in the inductive setting: new nodes.

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

Transductive Setting and MLP Sizes

Table 1: GLNNs outperform MLPs by large margins and match GNNs on 5 of 7 datasets under the **transductive** setting. Δ_{MLP} (Δ_{GNN}) represents difference between the GLNN and a trained MLP (GNN). Results show accuracy (higher is better); $\Delta_{GNN} \geq 0$ indicates GLNN outperforms GNN.

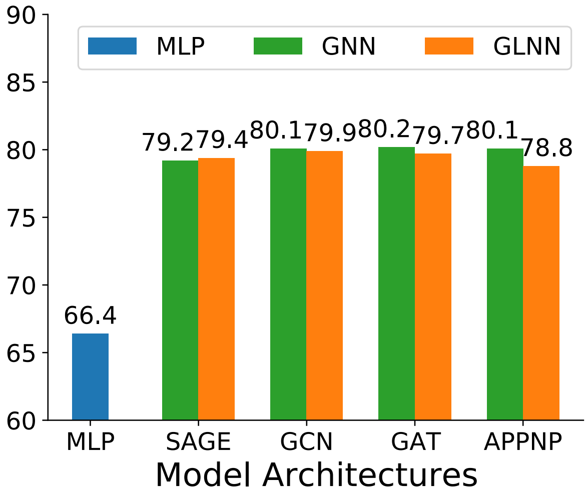
Datasets	SAGE	MLP	GLNN	Δ_{MLP}	Δ_{GNN}
Cora	80.52 ± 1.77	59.22 ± 1.31	$\mathbf{80.54 \pm 1.35}$	21.32 (36.00%)	0.02 (0.02%)
Citeseer	70.33 ± 1.97	59.61 ± 2.88	$\mathbf{71.77 \pm 2.01}$	12.16 (20.40%)	1.44 (2.05%)
Pubmed	75.39 ± 2.09	67.55 ± 2.31	$\mathbf{75.42 \pm 2.31}$	7.87 (11.65%)	0.03 (0.04%)
A-computer	82.97 ± 2.16	67.80 ± 1.06	$\mathbf{83.03 \pm 1.87}$	15.23 (22.46%)	0.06 (0.07%)
A-photo	90.90 ± 0.84	78.77 ± 1.74	$\mathbf{92.11 \pm 1.08}$	13.34 (16.94%)	1.21 (1.33%)
Arxiv	$\mathbf{70.92 \pm 0.17}$	56.05 ± 0.46	63.46 ± 0.45	7.41 (13.24%)	-7.46 (-10.52%)
Products	$\mathbf{78.61 \pm 0.49}$	62.47 ± 0.10	68.86 ± 0.46	6.39 (10.23%)	-9.75 (-12.4%)

Table 2: Enlarged GLNNs match the performance of GNNs on the OGB datasets. For `Arxiv`, we use MLPw4 (GLNNw4). For `Products`, we use MLPw8 (GLNNw8).

Datasets	SAGE	MLP+	GLNN+	Δ_{MLP}	Δ_{GNN}
Arxiv	70.92 ± 0.17	55.31 ± 0.47	$\mathbf{72.15 \pm 0.27}$	16.85 (30.46%)	0.51 (0.71%)
Products	$\mathbf{78.61 \pm 0.49}$	64.50 ± 0.45	77.65 ± 0.48	13.14 (20.38%)	-0.97 (-1.23%)

S. Zhang, Y. Liu, Y. Sun, N. Shah. Graph-less Neural Networks, ICLR 2022

GLNN with Different Teach GNNs



GLNN works with different GNN architectures as the teacher model

GLNN with One-hop Feature Augmentation

1. 1-hop GA-MLP: firstly, for each node v, we collect features of its 1-hop neighbors u to augment the raw feature of v, i.e. $x_v \rightarrow \tilde{x}_v$, like in SGC. Then we train an MLP on the graph with $\tilde{x}_v$. Note if v is in the observed graph but u is in the inductive (unobserved during training) part, then v doesn't collect features from u.
2. 1-hop GA-GLNN: Go through the same feature augmentation step as 1-hop GA-MLP. Then train an MLP with distillation from teacher GNN.
3. In summary, we compare 5 different models in the table below
   (a) SAGE: single model on x_v
   (b) MLP: single model on x_v
   (c) GLNN: SAGE teacher and MLP student on x_v
   (d) 1-hop GA-MLP: single model on $\tilde{x}_v$
   (e) 1-hop GA-GLNN: SAGE teacher on x_v, MLP student on $\tilde{x}_v$

	Eval	SAGE	MLP	GLNN	1-hop GA-MLP	1-hop GA-GLNN
`Arxiv`	*ind*	70.64	55.40	60.48	66.62	68.83
	tran	70.75	55.28	71.46	66.67	69.82