HOW ATTENTIVE ARE GRAPH ATTENTION NETWORKS?

Shaked Brody, Uri Alon, Eran Yahav
ICLR 2022

BACKGROUND

GRAPH NEURAL NETWORKS (GNNS)

- (GNNs) are central for learning on graph-structured data.

GRAPH ATTENTION NETWORK (GAT)

1. Graph Attention Network (GAT, Velickovic et al., 2018) is widely used for learning node representations via attention.

[1]Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4):18–42, 2017.
[2]Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veli˘ckovi´c. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

PROBLEM STATEMENT

01 - DYNAMIC ATTENTION

- GAT is believed to compute dynamic attention.Different queries can rank neighbors differently.

02 - STANDARD GAT COMPUTES ONLY STATIC ATTENTION

- The ranking of attention scores is fixed, unconditioned on the query node.

03 - WHY IS DYNAMIC ATTENTION IMPORTANT

- Enables modeling of contexts where different nodes attend to different neighbors.

Wu Z, Ying C, Zhao F, Fan Z, Dai X, Xia R. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. arXiv preprint arXiv:2010.04640. 2020 Oct 9.

Chen Z, Huang H, Liu B, Shi X, Jin H. Semantic and syntactic enhanced aspect sentiment triplet extraction. arXiv preprint arXiv:2106.03315. 2021 Jun 7.

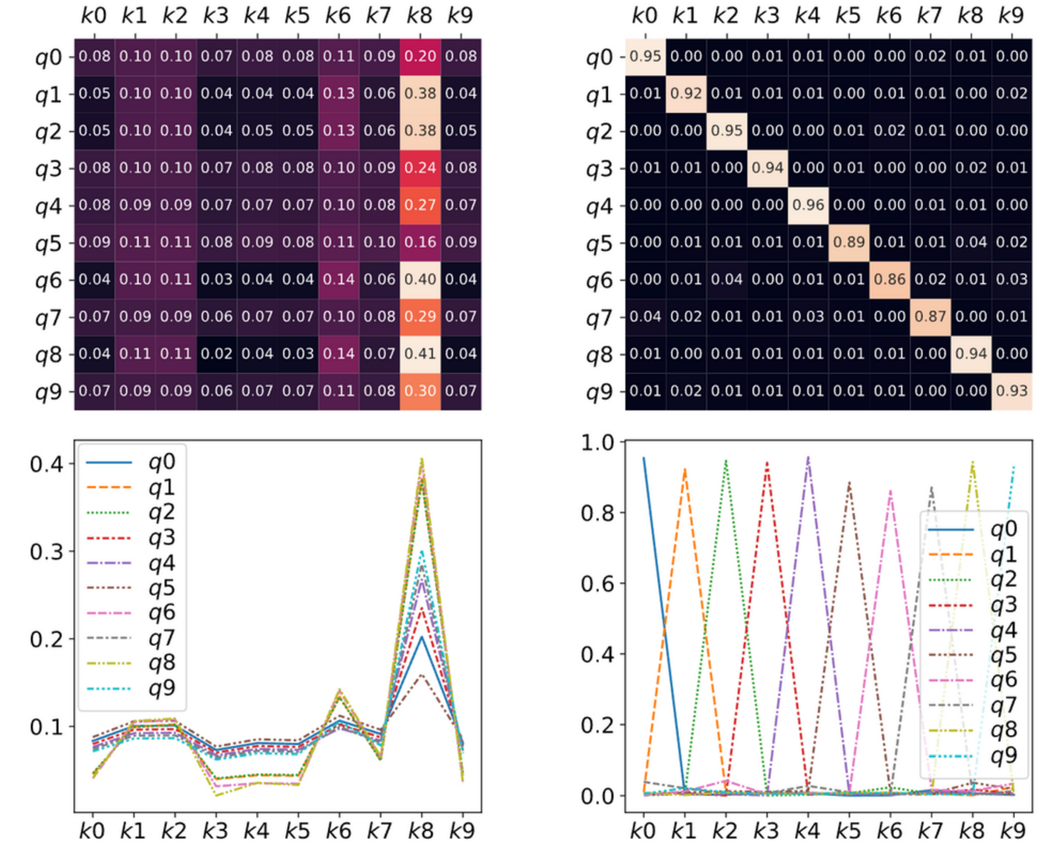
STATIC VS. DYNAMIC ATTENTION

01 - STATIC ATTENTION

- The attention coefficients for neighbors are shared across all queries (nodes); cannot select different neighbors per query.

02 - DYNAMIC ATTENTION

- The neighbor ranking can differ for each query node, as in traditional attention models..



(a) Attention in standard GAT (Veličković et al. (2018))

(b) Attention in GATv2, our fixed version of GAT

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010, 2017.

Petar Veli ˇ ckovi ´c, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In International Conference on Learning Representations, 2018.

Petar Veli˘ckovi ć, Lars Buesing, Matthew Overlan, Razvan Pascanu, Oriol Vinyals, and Charles Blundell. Pointer graph networks. Advances in Neural Information Processing Systems, 33, 2020. Petar et al. Veli˘ckovi ć. Graph attention networks. 2018.

THEORETICAL ANALYSIS

01 - WHAT DOES THIS MEAN MATHEMATICALLY?

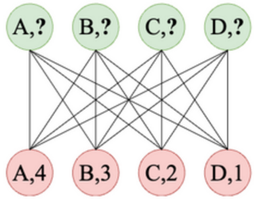
- In GAT, however, the attention score for a neighbor depends only on the neighbor's transformed features (the key), not on the specific query node.

02 - THEOREM 1

- For any set of node features, GAT's attention scores can be sorted once and that order is shared across all nodes in the graph

03 - WHY IS THIS A PROBLEM?

- Static attention limits expressivity



GATV2 – THE SOLUTION

SIMPLE CHANGE

- Switch order of linear and nonlinearity operations in the attention scoring function

THEOREM 2

- GATv2 enables dynamic attention and is a universal approximator for attention functions

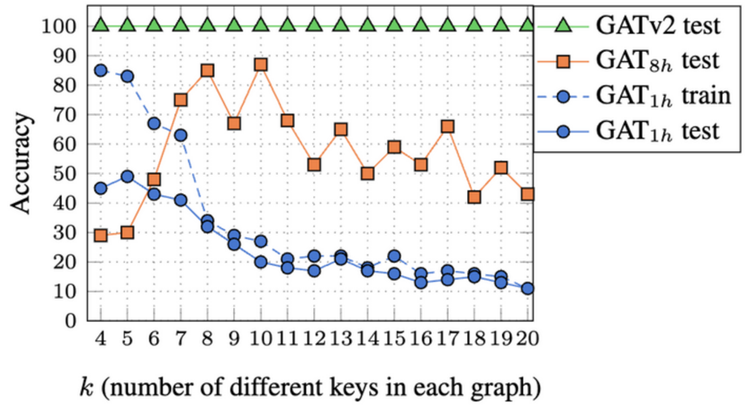
Table 1: Average accuracy (Table 1a) and ROC-AUC (Table 1b) in node-prediction datasets (10 runs±std). In all datasets, GATv2 outperforms GAT. † – previously reported by Hu et al. (2020).

		(a)			(b)
Model	Attn. Heads	**ogbn-arxiv**	**ogbn-products**	**ogbn-mag**	**ogbn-proteins**
GCN†	0	71.74±0.29	78.97±0.33	30.43±0.25	72.51±0.35
GraphSAGE†	0	71.49±0.27	78.70±0.36	31.53±0.15	77.68±0.20
GAT	1	71.59±0.38	79.04±1.54	32.20±1.46	70.77±5.79
	8	71.54±0.30	77.23±2.37	31.75±1.60	78.63±1.62
GATv2 (this work)	1	71.78±0.18	**80.63**±0.70	**32.61**±0.44	77.23±3.32
	8	**71.87**±0.25	78.46±2.45	32.52±0.39	**79.52**±0.55

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. arXiv preprint arXiv:2005.00687, 2020.

SYNTHETIC BENCHMARK

01 - DICTIONARY LOOKUP: REQUIRES DYNAMIC ATTENTION TO SOLVE.



02 - GAT FAILS TO FIT TRAINING DATA EVEN WITH MANY HEADS.

03 - GATV2 SUCCEEDS WITH PERFECT ACCURACY AND GENERALIZATION

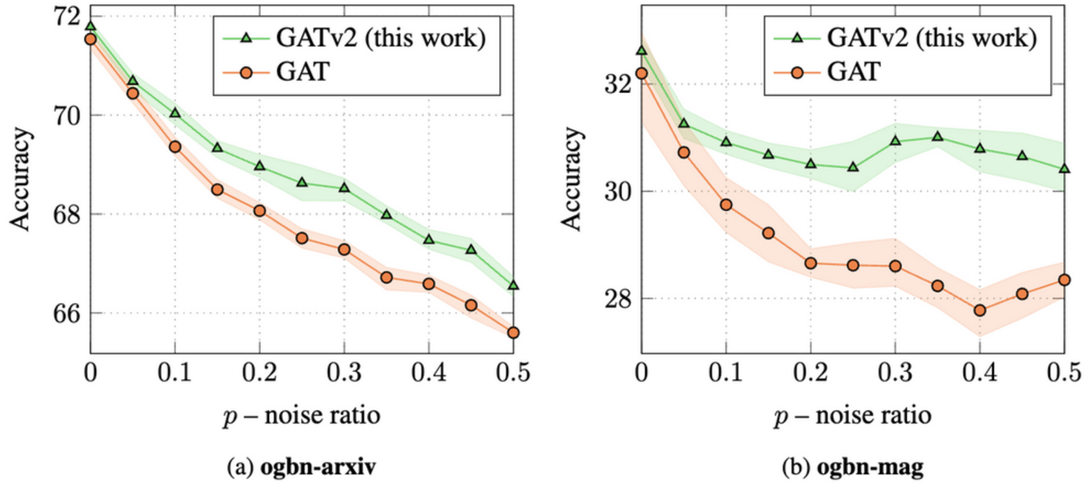
EXPERIMENT HIGHLIGHTS

01 - ROBUSTNESS TO EDGE NOISE

- GATv2 more resistant than GAT to structural noise in test graphs.

02 - BENCHMARKS

- (OGB node, link, graph prediction): GATv2 consistently outperforms GAT in accuracy and robustness..
- Example: On the QM9 dataset, GATv2 achieves 11.5% lower error than GAT across 13 quantum property prediction tasks.



(a) **ogbn-arxiv** (b) **ogbn-mag**

Figure 4: Test accuracy compared to the noise ratio: GATv2 is more robust to structural noise compared to GAT. Each point is an average of 10 runs, error bars show standard deviation.

Table 2: Average error rates (lower is better), 5 runs for each property, on the QM9 dataset. The best result among GAT and GATv2 is marked in **bold**; the globally best result among all GNNs is marked in **bold and underline**. † was previously tuned and reported by Brockschmidt (2020).

Model	Predicted Property 1	2	3	4	5	6	7	8	9	10	11	12	13	Rel. to GAT
GCN†	3.21	**4.22**	1.45	1.62	2.42	16.38	17.40	7.82	8.24	9.05	7.00	3.93	**1.02**	-1.5%
GIN†	**2.64**	4.67	1.42	1.50	**2.27**	**15.63**	**12.93**	**5.88**	18.71	**5.62**	**5.38**	**3.53**	1.05	-2.3%
GAT†	2.68	4.65	1.48	1.53	2.31	52.39	14.87	7.61	6.86	7.64	6.54	4.11	**1.48**	+0%
GATv2	**2.65**	**4.28**	**1.41**	**1.47**	**2.29**	**16.37**	**14.03**	**6.07**	**6.28**	**6.60**	**5.97**	**3.57**	1.59	**-11.5%**

DISCUSSION

01 - GAT'S POPULARITY IS BASED ON A MISCONCEPTION OF ITS TRUE ATTENTION MECHANISM.

02 - GATV2 LIFTS LIMITATIONS

- should be preferred for problems requiring expressive, query-conditioned attention.

CONCLUSION

- Identification of static attention in GAT; introduction and demonstration of GATv2.
- GATv2 is strictly more expressive and robust; minimal overhead compared to GAT.

code is
available at https://github.com/tech-srl/how_attentive_are_gats.1

THANK YOU