How Attentive are Graph Attention Networks?

Shaked Brody
Technion, Israel
shakedbr@cs.technion.ac.il

Uri Alon
Technion, Israel
urialon@cs.technion.ac.il

Eran Yahav
Technion, Israel
yahave@cs.technion.ac.il

GRAPH ATTENTION NETWORKS

Petar Veličković*
Department of Computer Science and Technology
University of Cambridge
petar.velickovic@cst.cam.ac.uk

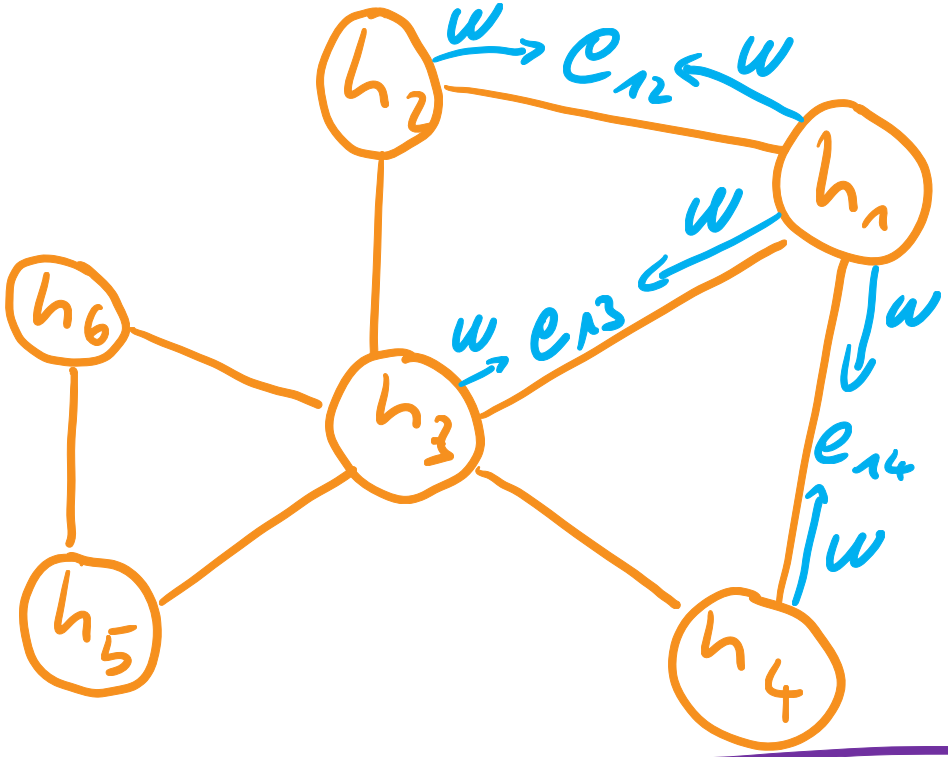
Guillem Cucurull*
Centre de Visió per Computador, UAB
gcucurull@gmail.com

Arantxa Casanova*
Centre de Visió per Computador, UAB
ar.casanova.8@gmail.com

Adriana Romero
Montréal Institute for Learning Algorithms
adriana.romero.soriano@umontreal.ca

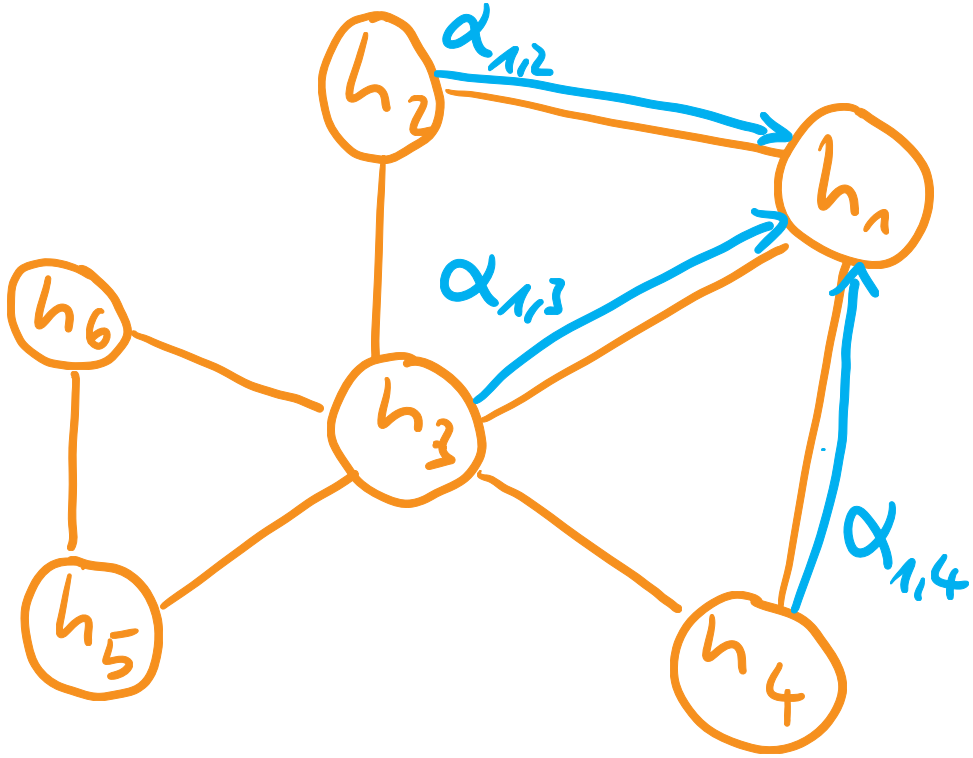
Pietro Liò
Department of Computer Science and Technology
University of Cambridge
pietro.lio@cst.cam.ac.uk

Yoshua Bengio
Montréal Institute for Learning Algorithms
yoshua.umontreal@gmail.com



1. coefficients

$$e_{ij} = \text{LeakyReLU}\left(\boldsymbol{a}^\top \cdot [\boldsymbol{W}\boldsymbol{h}_i \| \boldsymbol{W}\boldsymbol{h}_j]\right)$$

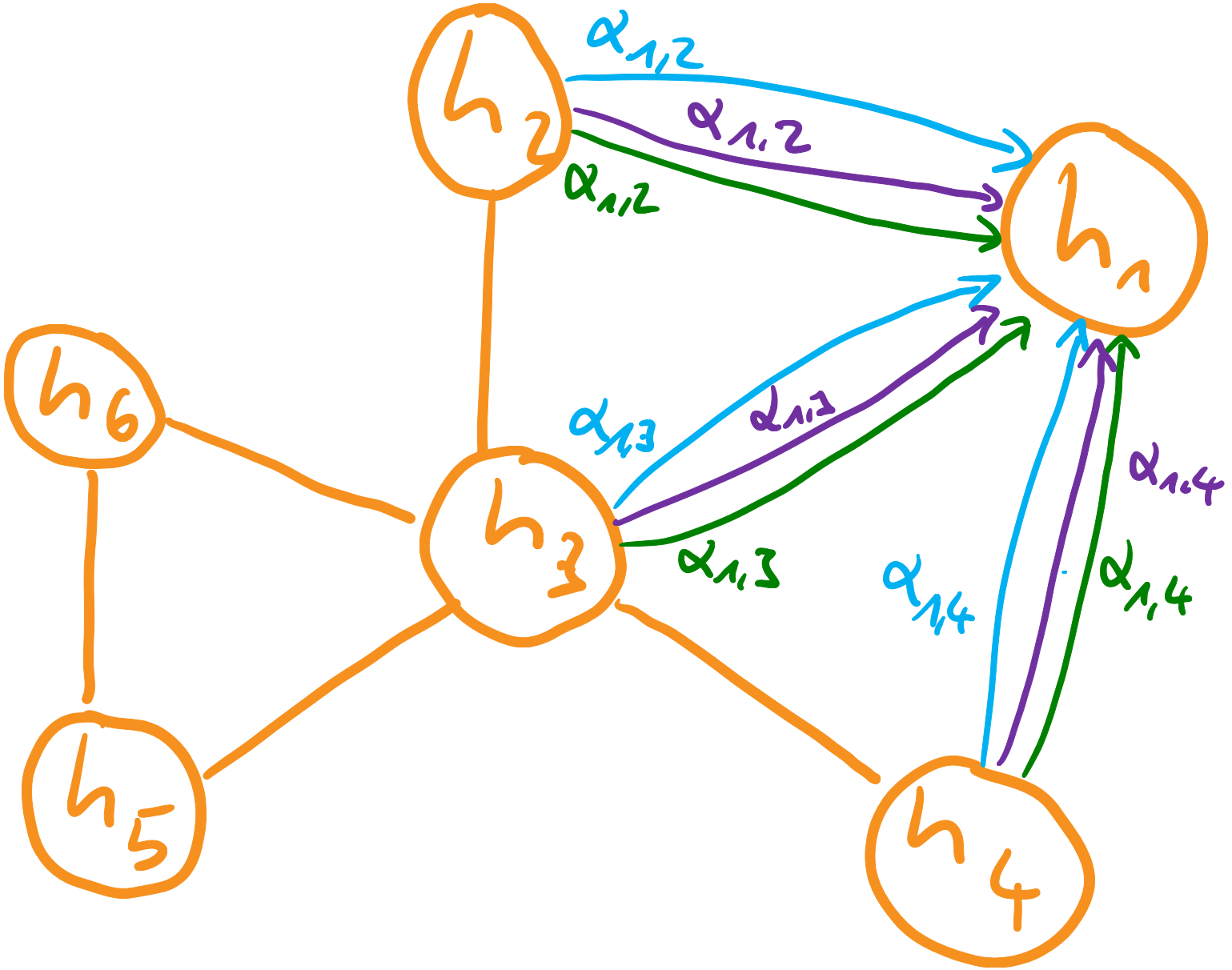


2. normalize:

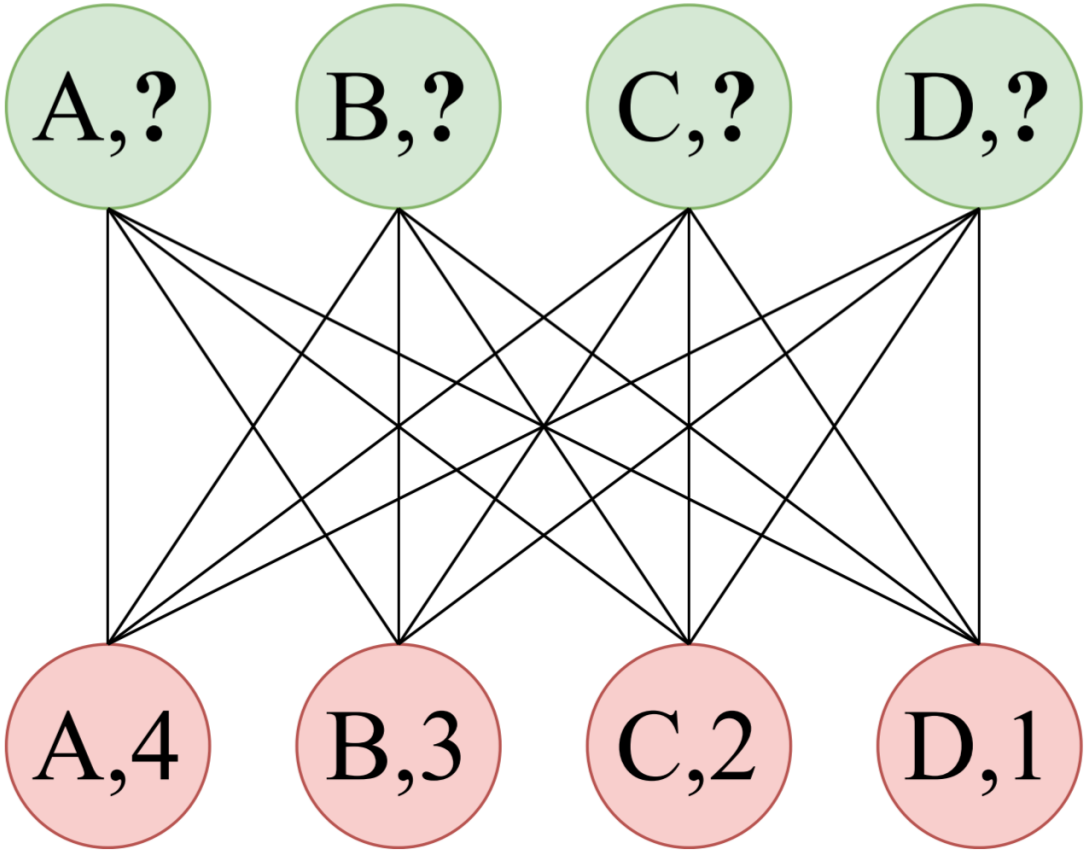
$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}.$$

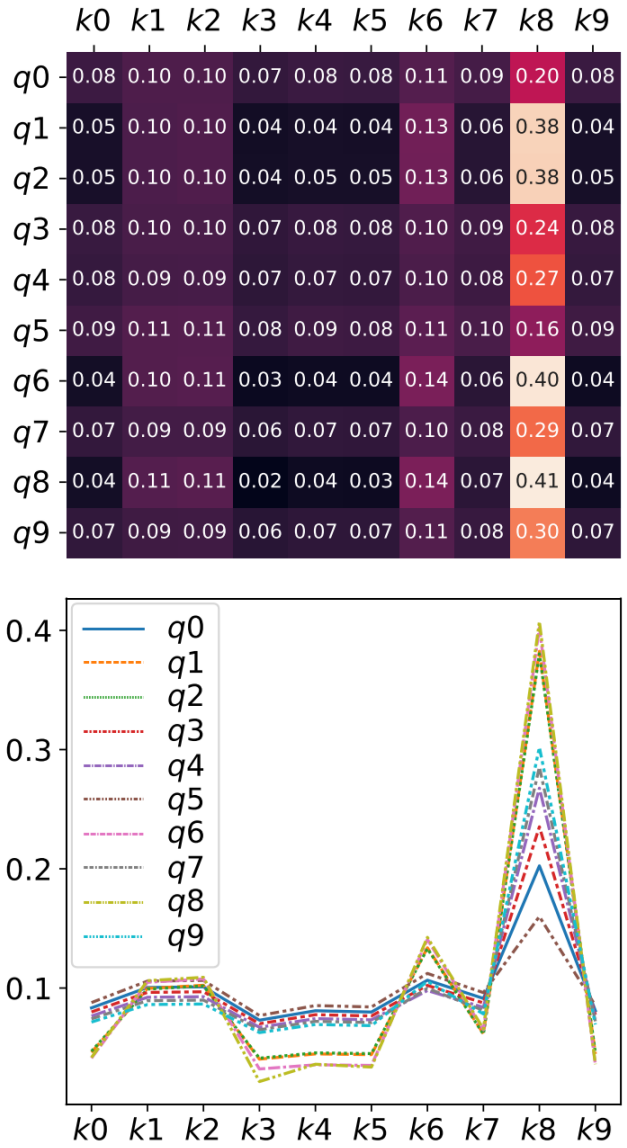
3. aggregate

$$\boldsymbol{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \boldsymbol{W}\boldsymbol{h}_j\right)$$

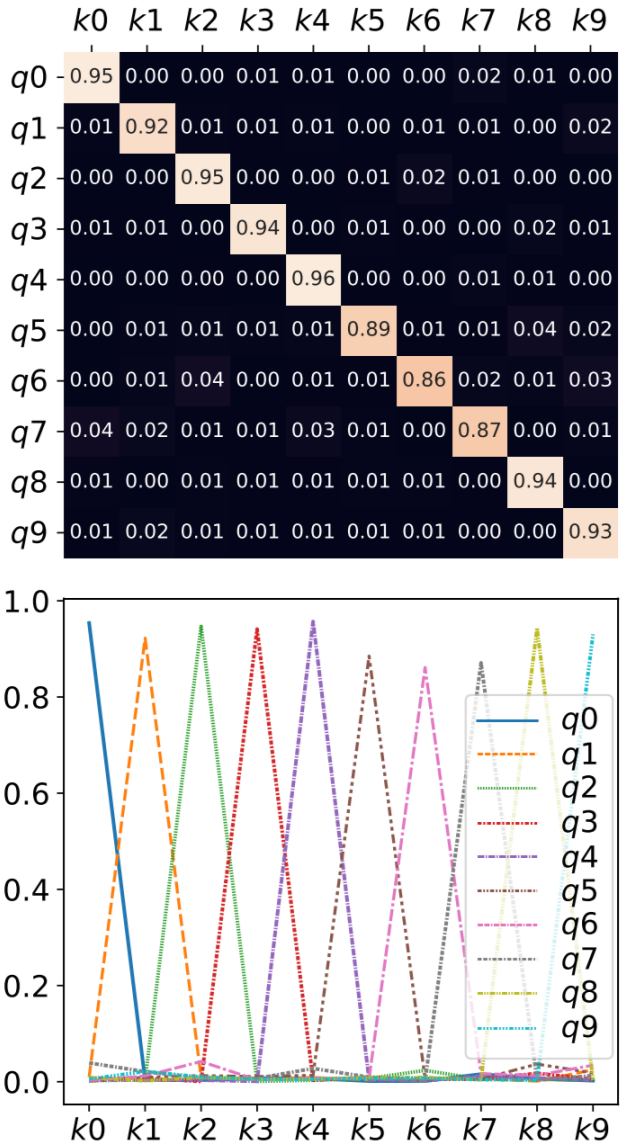

h_2
h_1
h_6
h_3
h_5
h_4
$\alpha_{1,2}$
$\alpha_{1,2}$
$\alpha_{1,2}$
$\alpha_{1,3}$
$\alpha_{1,3}$
$\alpha_{1,3}$
$\alpha_{1,4}$
$\alpha_{1,4}$
$\alpha_{1,4}$

$h'_1 \| h'_1 \| h'_1$


A,?
B,?
C,?
D,?
A,4
B,3
C,2
D,1



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



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

$$e_{i,j} = \text{LeakyReLU}\left(a^T \cdot Wh_i \,\|\, Wh_j\right)$$

$$e_{i,j} = \left(\boxed{a^T} \cdot \begin{bmatrix} W \cdot h_i \\ W \cdot h_j \end{bmatrix}\right)$$

$$e_{i,j} = \left(\boxed{a_1^T}\,\boxed{W \cdot h_i} + \boxed{a_2^T}\,\boxed{W \cdot h_j}\right)$$

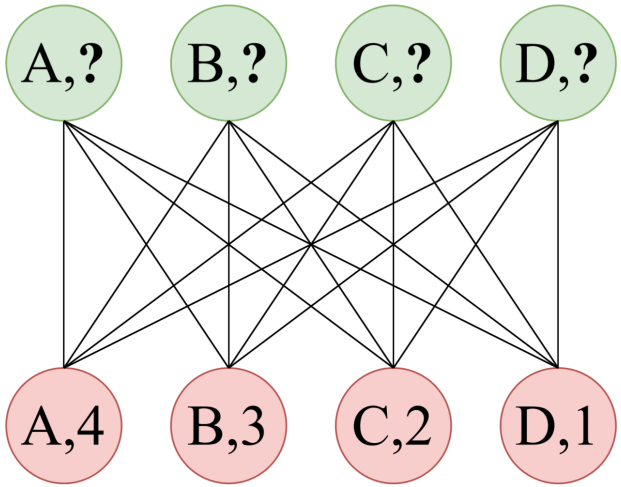
$$e_{i,j} = \text{LeakyReLU}\left(a_1^T \cdot Wh_i + a_2^T \cdot Wh_j\right)$$

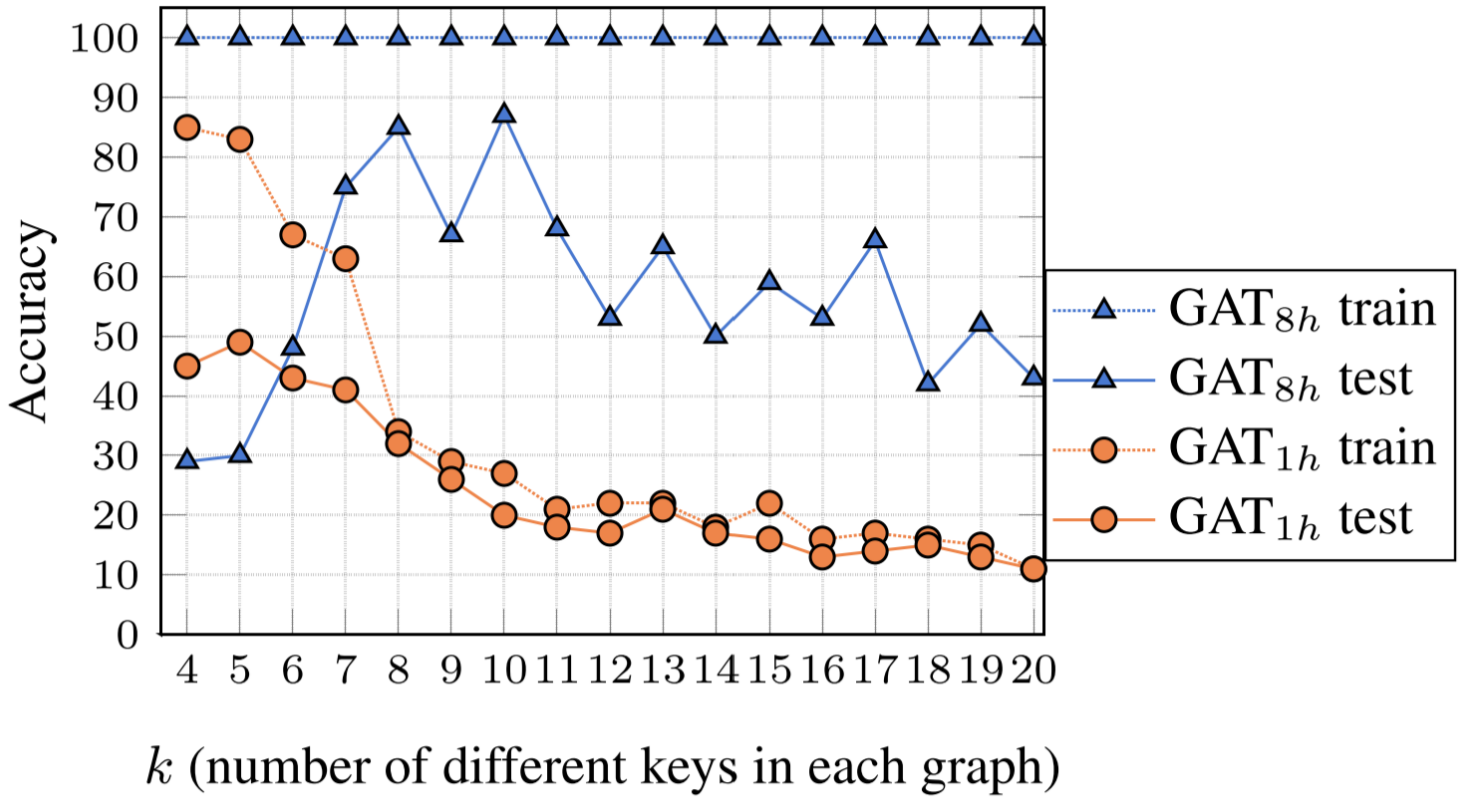
$$e_{ij} = \text{LeakyReLU}\left(a^T \cdot Wh_i \| Wh_j\right)$$

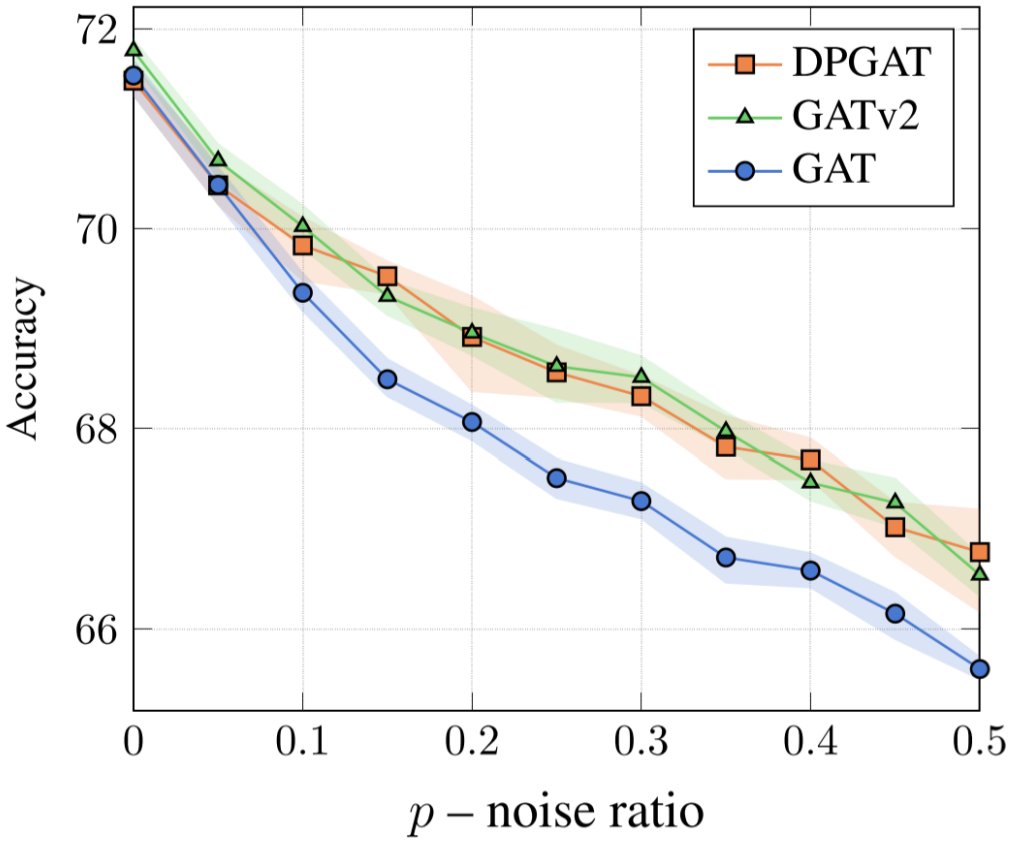
$$e_{ij} = \left(\boxed{a^T} \cdot \begin{matrix} \boxed{W} \cdot \boxed{h_i} \\ \boxed{W} \cdot \boxed{h_j} \end{matrix}\right)$$

$$e_{ij} = a^T \cdot \text{LeakyReLU}\left(W\, h_i \| h_j\right)$$

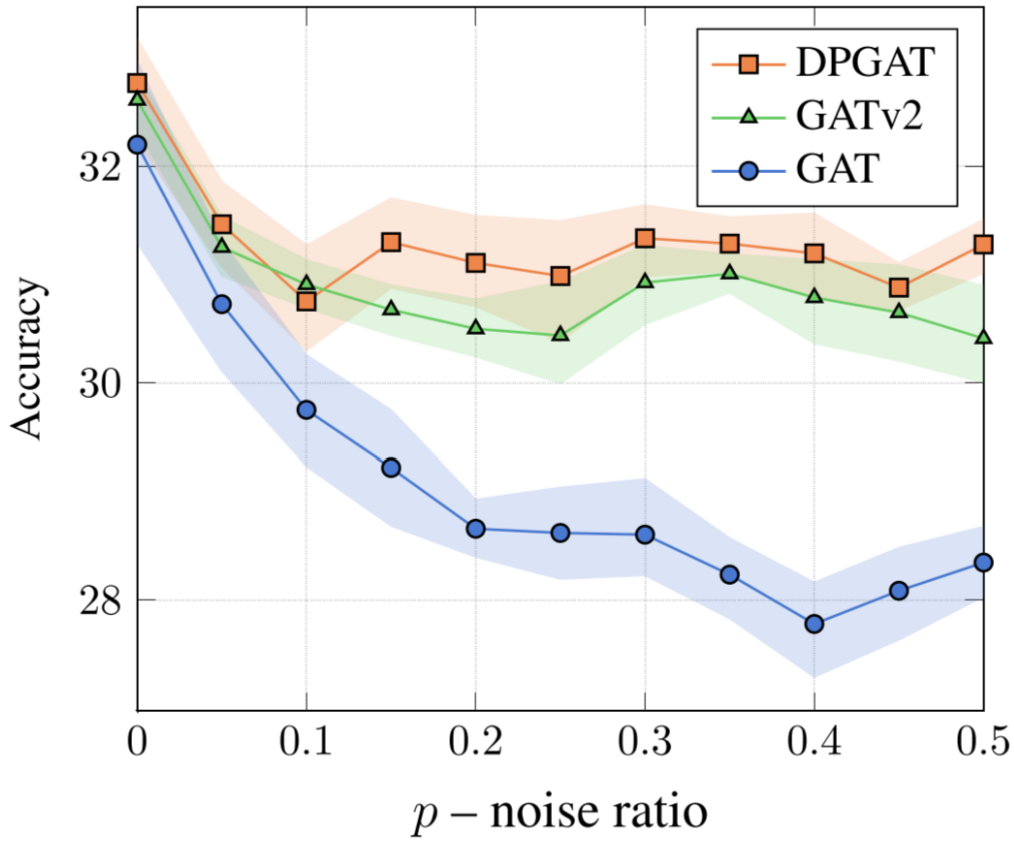
$$e_{ij} = \boxed{a^T} \cdot \left(\boxed{W} \begin{matrix} \boxed{h_i} \\ \boxed{h_j} \end{matrix}\right)$$


A,?
B,?
C,?
D,?
A,4
B,3
C,2
D,1

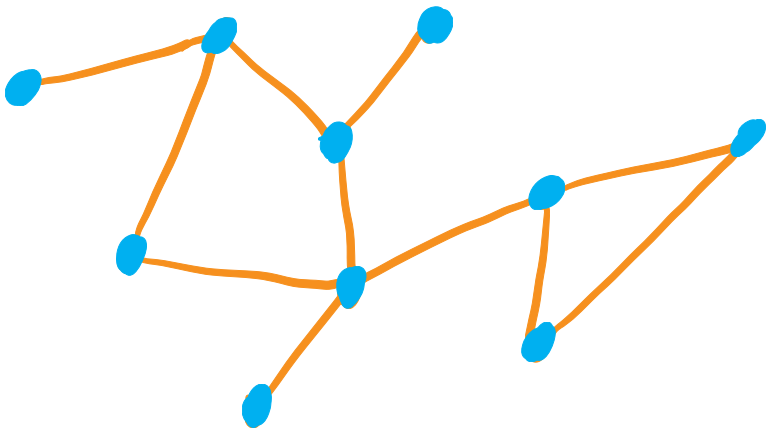

Accuracy
100
90
80
70
60
50
40
30
20
10
0
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
k (number of different keys in each graph)
GAT8h train
GAT8h test
GAT1h train
GAT1h test



(a) **ogbn-arxiv**



(b) **ogbn-mag**



	Model	**ogbn-arxiv**	**ogbn-products**	**ogbn-mag**
No-Attention	GCN†	71.74±0.29	78.97±0.33	30.43±0.25
	GraphSAGE†	71.49±0.27	78.70±0.36	31.53±0.15
Attention 1 head	GAT_{1h}	71.59±0.38	79.04±1.54	32.20±1.46
	$DPGAT_{1h}$	71.52±0.17	76.49±0.78	**32.77**±0.80
	$GATv2_{1h}$	71.78±0.18	**80.63**±0.70	32.61±0.44
Attention 8 heads	GAT_{8h}	71.54±0.30	77.23±2.37	31.75±1.60
	$DPGAT_{8h}$	71.48±0.26	73.53±0.47	27.74±9.97
	$GATv2_{8h}$	**71.87**±0.25	78.46±2.45	32.52±0.39

(a)

ogbn-proteins
72.51±0.35
77.68±0.20
70.77±5.79
63.47±2.79
77.23±3.32
78.63±1.62
72.88±0.59
79.52±0.55

(b)

Table 2: Average accuracy (Table 2a) and ROC-AUC (Table 2b) in node-prediction datasets (10 runs±std). In all datasets, GATv2 outperforms GAT. † – previously reported by Hu et al. (2020).

	Model	ogbl-collab w/o val edges	ogbl-collab w/ val edges
No-Attention	GCN†	44.75±1.07	47.14±1.45
	GraphSAGE†	48.10±0.81	54.63±1.12
Attention 1 head	GAT_{1h}	39.32±3.26	48.10±4.80
	$DPGAT_{1h}$	44.15±1.59	49.61±3.16
	$GATv2_{1h}$	42.00±2.40	48.02±2.77
Attention 8 heads	GAT_{8h}	42.37±2.99	46.63±2.80
	$DPGAT_{8h}$	**48.99**±0.70	**56.90**±1.45
	$GATv2_{8h}$	42.85±2.64	49.70±3.08

(a)

ogbl-citation2
80.04±0.25
80.44±0.10
79.84±0.19
80.31±0.21
80.33±0.13
75.95±1.31
79.64±0.42
80.14±0.71

(b)

Table 3: Average Hits@50 (Table 3a) and mean reciprocal rank (MRR) (Table 3b) in link-prediction benchmarks from OGB (10 runs±std). † – previously reported by Hu et al. (2020).

Model	Predicted Property 1	2	3	4	5	6	7	8	9	10	11	12	13	Rel. to GAT_{8h}
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_{1h}	3.08	7.82	1.79	3.96	3.58	35.43	116.5	28.10	20.80	15.80	10.80	5.37	3.11	+134.1%
$DPGAT_{1h}$	3.20	8.35	1.71	2.17	2.88	25.21	65.79	12.93	13.32	14.42	13.83	6.37	3.28	+77.9%
$GATv2_{1h}$	3.04	6.38	1.68	2.18	2.82	20.56	77.13	10.19	22.56	15.04	22.94	5.23	2.46	+91.6%
GAT_{8h}†	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%
$DPGAT_{8h}$	**2.63**	4.37	1.44	**1.40**	**2.10**	32.59	**11.66**	6.95	7.09	7.30	6.52	3.76	1.18	-9.7%
$GATv2_{8h}$	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%**

Table 4: Average error rates (lower is better), 5 runs for each property, on the QM9 dataset. † was previously tuned and reported by Brockschmidt (2020). Standard deviation is reported in Appendix F.

$$e_{i,j} = \frac{h_i^T Q \cdot (h^T K)^T}{\sqrt{d_k}}$$