Multi-Stain Self-Attention Graph Multiple Instance Learning Pipeline for Histopathology Whole Slide Images

Amaya Gallagher-Syed[1, 2]
a.r.syed@qmul.ac.uk
Luca Rossi[3]
luca.rossi@polyu.edu.hk
Felice Rivellese[2]
f.rivellese@qmul.ac.uk
Costantino Pitzalis[2]
c.pitzalis@qmul.ac.uk
Myles Lewis[2]
myles.lewis@qmul.ac.uk
Michael Barnes[1,2]
m.r.barnes@qmul.ac.uk
Gregory Slabaugh[2]
g.slabaugh@qmul.ac.uk

[1] Digital Environment Research Institute
Queen Mary University of London
London, UK

[2] William Harvey Research Institute
Queen Mary University of London
London, UK

[3] Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hong Kong

1 Supplementary material

1.1 Memory Usage

In Figure 1 we show the increase in GPU RAM [GB] cost and FLOPs [M] in a forward pass for different values of connectivity K in a k-NNG with 2000 nodes, approximately the average graph size for the dataset. Despite scaling linearly in the observed range, said k-NNG with a $k = 100$ already requires 4 GB RAM during the forward pass. This can quickly become prohibitive for any work done on a commercial workstation, as GPUs have limited on-chip memory [7]. This highlights the need to use the minimally connected graph, so high memory needs don't prevent users from performing training and inference. This value k is a parameter of the model and requires initial hyperparameter tuning to choose an optimal value.

1.2 Further model ablation

We conduct further ablation on the model, to disaggregate the contribution of each part. We run the GAT_SAGPool model with between 1 to 5 layers and we look at the number of attention heads, going from 1 to 8.

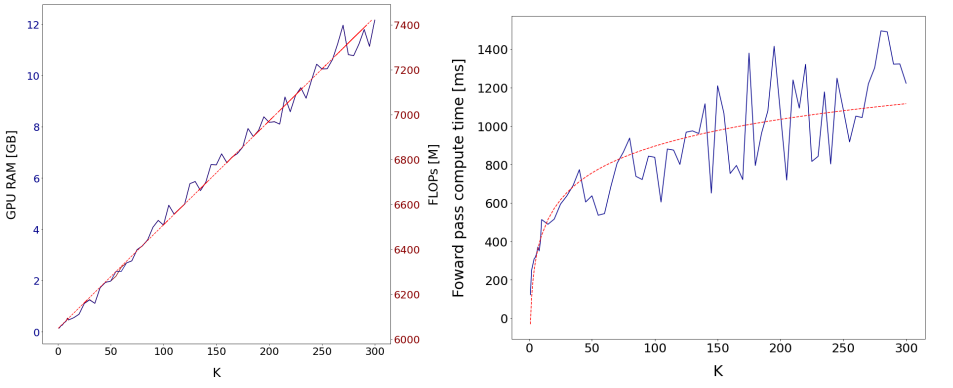

Figure 1: *Memory and resource usage per k in a 2000 nodes k-NNG for a model forward pass. Left: GPU RAM [GB] usage. Right: Compute time in ms*

Number of layers In Table 1 we show the increase in accuracy with each additional GAT + SAGPool layer. We note that 5 layers also obtain best F1-score, however at a higher computational cost.

	GAT_SAGPool
1 layer	0.80
2 layer	0.83
3 layer	0.87
4 layer	0.89
5 layer	0.89

Table 1: F1-score for a GAT_SAGPool model with 1 to 5 layers

Number of heads From looking at the results in Table 2 we see 2 or 8 averaged attention heads obtain best F1-score, yet both 1 and 4 also heads obtain close results. Because this indicates the model does not need many attention heads for optimal performance, using 2 attention heads seems sufficient [1], however this is a hyperparameter that can be chosen by the user based on the specifics of their dataset.

	F1-score
1 head	0.87
2 heads	0.89
4 heads	0.87
8 heads	0.89

Table 2: F1-score for different numbers of averaged attention heads.

References

[1] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. Multi-Head Attention: Collaborate Instead of Concatenate. Technical report, May 2021. URL http://arxiv.org/abs/2006.16362. arXiv:2006.16362 [cs, stat] type: article.

[2] Zhaokang Wang, Yunpan Wang, Chunfeng Yuan, Rong Gu, and Yihua Huang. Empirical analysis of performance bottlenecks in graph neural network training and inference with GPUs. *Neurocomputing*, 446:165–191, July 2021. ISSN 0925-2312. doi: 10.1016/j.neucom.2021.03.015. URL https://www.sciencedirect.com/science/article/pii/S0925231221003659.