Improving whole-slide images analysis via intra- and inter-scale random graph neural networks

Ali Abdari[1,2], Sonali Andani[3], Roberto Basla[4], Giacomo Capitani[5], Alex Falcon[1], Putra Manggala[6]
1: Uni. Udine; 2: UniNA Federico II; 3: ETH Zurich; 4: PoliMI; 5: UniMORE; 6: Uni. Amsterdam

1. Introduction

The digitalization of pathology images supports the work of clinicians by enabling automation, increasing analysis quality, and reducing their workload. For example, traditional glass slides acquired by microscopes can now be digitized into *whole-slide images (WSIs)*. WSIs are pyramidal images with access to global tissue features at a low resolution and cell level morphological features at a high resolution (Figure 1). They are used as the gold standard information source for cancer diagnosis. Deep learning can aid in their classification by identifying and segmenting tissues containing tumors [2,4,5].

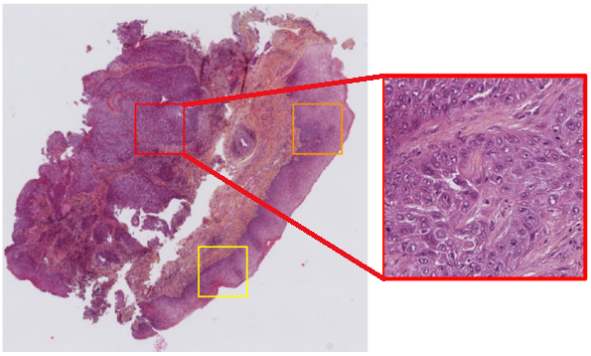

Figure 1: Example of a whole-slide image depicting the global and zoomed-in tissue features (adapted from [1]).

WSIs can be as large as 100K×100K pixels, and thus cannot be processed by a neural network with a reasonable amount of memory and time. Modern multi-instance learning (MIL) processes each slide image as a bag composed of a set of patches (instances) and aggregates them to produce bag-level classification via an attention mechanism [2,6]. A recent approach, termed DAS-MIL [2], leverages the pyramidal-structured WSIs by first performing self-supervision to extract instance features at different scales, followed by learning a deep learning model to perform a bag-level classification. DAS-MIL proposes the use of Graph Attention Networks (GAT) [7] to perform representation learning on the multi-scale WSI patches, by setting the following fixed choice of graph neighborhood structure:

Each node is connected to the nearest 8 patches within the same scale and lower resolution patches are connected to higher resolution patches they contain.

Features extracted by GAT are then inputted into a MIL model to obtain bag-level classification via aggregating instance-level prediction.

While existing graph neural networks typically include all neighbors in the same scope for feature aggregation [8], a recent study has shown that adapting each neighborhood to attend to more relevant neighbors can enable GAT to learn better representations [3]. This work proposes various mechanisms for downweighting irrelevant neighbors during attention computation. Following the same line of thought, we examine whether the fixed choice of graph neighborhood in [2] induces the most relevant neighbors that would allow for optimal GAT representation learning.

Our contributions are:

1. We contribute quality of life (QoL) improvements based on machine learning engineering best practices: the use of repeated metric measurements and three-way-splits. We elaborate how our QoL improvements allow for a more reliable and trustworthy metric measurement.
2. We empirically validate that random intra- and inter-scales selection of edges is a more performant design choice than the fixed neighborhood structure proposed in [2]. This shows that the fixed choice in [3] does not induce the most relevant neighbors for GAT aggregation and a random graph family can instead be chosen to parametrize the neighborhood structure for GAT.

2. Methodology

2.1 Quality of life (QoL) improvements: fixing data leakage and increasing the approximation quality of generalization performance

We find that the implementation of [3] makes use of a two-way split, training and test, instead of a three-way, training, validation and test. This is due to the small number of WSIs under analysis. For instance, Camelyon16 has 270 WSIs for training. However, for obtaining the best performing model weights, DAS-MIL relies on the test set. This test data leakage can cause the trained model to not generalize well in practice [9]. We extract ten percent of the training set to create a validation set. We use this validation set for best model selection.

For every proposed method, we perform a full training-validation-test loop ten times to obtain a better approximation of generalization performance.

2.2 Random graph attention networks

Motivated by the observations in [3], we examine whether the graph structure proposed in [2] (Figure 2.a) for DAS-MIL induces the most relevant neighbors for attention aggregation. We address this by choosing random edges between nodes/patches intra- and inter-scales instead of based on nearest neighbor (Figure 2.b). We hypothesize that replacing nearest neighbor edges with randomly selected edges can induce more relevant neighbors for an improved information flow across the WSI. We quantify this relevance by measuring predictive performance for tumor classification.

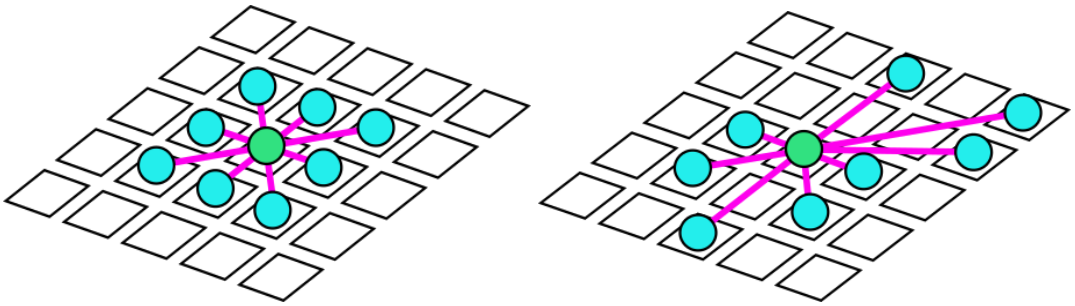

Figure 2: (a) Original graphs. (b) Random graph (E=8).

3. Experiments

Similar to in [2], we use the Camelyon16 dataset and report the average accuracy and AUC (over 10 runs with different seeds).

3.1 QoL improvements

We investigate the impact of performing and summarizing the results from repeated runs and three-way-split (See Section 2.1 for definition). The results are summarized In Table 1.

We first observe that for the two-split runs, the test performance metric values obtained from single runs (DAS-MIL results from [2]) are outlying with respect to the mean and standard deviations we have computed (DAS-MIL from our QoL-improved run). We also observe that our QoL-improved run with a three-way-split yields lower test performance metric values. The metric values reported by the two-way-split are likely to be inflated and are an overestimate of the true generalization performance.

These results indicate the importance of summarizing from repeated runs and three-way-split in a machine learning pipeline. We propose the use of our DAS-MIL metric values instead of the ones reported in [2] when evaluating the model proposed in that paper.

Model	Camelyon16 (train-test)		Camelyon16 (train-val-test)	
	Accuracy	AUC (Lower)	Accuracy	AUC (Lower)
DAS-MIL results from [2]	0.945	0.973	—	—
DAS-MIL from our QoL-improved run	0.909 ± 0.004	0.947 ± 0.003	0.897 ± 0.004	0.942 ± 0.003

Table 1: Results obtained by running the baseline using the two different splits under analysis (train-test, and train-val-test).

3.2 Random graph attention networks

We empirically compare DAS-MIL with fixed neighborhood structure [2] and DAS-MIL with random edges. All test set performance metric values are computed using the QoL improvements presented in Section 3.1.

Model		Camelyon16	
		Accuracy	AUC
DAS-MIL		0.897 ± 0.004	0.942 ± 0.003
DAS-MIL with random edges (Ours)	E=1	0.901 ± 0.012	0.918 ± 0.017
	E=2	**0.917 ± 0.015**	**0.957 ± 0.011**
	E=4	0.898 ± 0.011	**0.958 ± 0.003**
	E=8	**0.908 ± 0.005**	0.939 ± 0.003
	E=16	0.902 ± 0.007	0.943 ± 0.006
	E=32	0.892 ± 0.010	0.942 ± 0.005

Table 2: Results obtained by running DAS-MIL with various numbers of random edges (E). The top-2 results for each metric are in bold.

We observe that DAS-MIL with random edges generally outperforms DAS-MIL, except for when E=1 and E=32. The top performing random models tend to use fewer edges than the fixed selection in [2], with E=2 and E=4. This shows that the random edges are able to induce more relevant patches for GAT aggregation.

4. Conclusions

When performing graph attention computation, it is vital that the relevant nodes are connected. We formulated a random graph variation of DAS-MIL that outperforms the fixed nearest-neighbor selection. For future WSI classification work, we conjecture that the choice of neighborhood should be treated either as an inferred latent variable (similar to [10]) or estimated as a hyperparameter. We posit that DAS-MIL with random graph neighborhood selection can be a strong and simple baseline for these more complicated learnable alternatives.

5. Acknowledgements

The Leonardo cluster proved to be instrumental by allowing us to test many ideas rapidly and reliably via repeated runs.

Bibliography

[1] Lubrano, M. *et al.*. Simple and Efficient Confidence Score for Grading Whole Slide Images. *arXiv:2303.04604* (2023).
[2] Bontempo, G. *et al*. DAS-MIL: Distilling Across Scales for MILClassification of Histological WSIs. In *MICCAI* (2023).
[3] Not All Neighbors Are Worth Attending to: Graph Selective Attention Networks for Semi-supervised Learning. He, T., Zhou, H., Ong, Y. S., Cong G. *arXiv:2210.07715* (2022).
[4] Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. Li, B., Li, Y., Eliceiri, K.W. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14318–14328 (2021).
[5] H2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L. In Proceedings of the AAAI Conference on Artificial Intelligence (2022).
[6] Attention-based Deep Multiple Instance Learning. Ilse, M., Tomczak, J., Welling, M. In International Conference on Machine Learning. vol. 80, pp. 2127–2136. PMLR (2018)
[7] Graph Attention Networks. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y. In International Conference on Learning Representations (2018)
[8] Decoupling the depth and scope of graph neural networks. Hanqing Zeng, Muhan Zhang, Yinglong Xia, Ajitesh Srivastava, Andrey Malevich, Rajgopal Kannan, Viktor Prasanna, Long Jin, and Ren Chen. In NeurIPS, Vol. 34. 19665–19679. (2021)
[9] Data wrangling and data leakage in machine learning for healthcare. Saravanan. N., Sathish. G.. In JETIR (2018)
[10] Bag Graph: Multiple Instance Learning using Bayesian Graph Neural Networks. Pal. S., et. al. *arXiv:2202.11132* (2022).