Comments for Reviewer wvfJ

Qiehe Sun

March 18, 2024

1 Comments

Thank you for recognizing our work and for your valuable suggestions. And we hope the following statements address your concerns::

- **Differences between MIL-RNN[CHG+19] and our work.** Our work differs significantly from MIL-RNN. Firstly, our approach stems from addressing informational redundancies within embedding-level methods and aims to improve them. In contrast, MIL-RNN primarily focuses on instance-level methods, where techniques like RNN and random forest serve as improvements over basic aggregation functions. Secondly, while MIL-RNN primarily deals with binary classification tasks in pathology, the BgIM dataset we collected is quadruple classified. Thirdly, although both our work and MIL-RNN involve pre-training the feature aggregator, a key distinction lies in strategy during training. MIL-RNN randomly samples instances within a bag, whereas we utilize all instances, leading to a longer inference but avoiding information loss. Fourthly, our aggregator structure represents a significant innovation in itself. Finally, in terms of instance filtering, MIL-RNN selects an equal number of discriminative instances as the number of recurrences, which is influenced by the mechanism of RNN. In contrast, we adopt a different approach by filtering a larger number of non-focal instances to create bag embeddings through bidirectional sampling. To illustrate this, based on your suggestion, we appended the MIL-RNN results on both datasets as shown in the table below. And NcIEMIL proved to be more effective.

Method	CAMELYON16			BgIM		
	ACC	AUC	F1-Score	ACC	AUC	F1-Score
MIL-RNN	$82.64_{2.17}$	$85.02_{2.73}$	$80.64_{1.94}$	$53.81_{3.56}$	$80.19_{1.86}$	$48.55_{2.81}$
NcIEMIL	$\mathbf{86.05_{1.55}}$	$\mathbf{89.68_{2.10}}$	$\mathbf{85.26_{1.54}}$	$\mathbf{85.23_{0.95}}$	$\mathbf{95.87_{0.60}}$	$\mathbf{81.20_{0.94}}$

- **Significance verification.** We reported the p-values of two-sample t-tests on ACC, AUC, and F1-Score between our method and baselines respectively. $p < 0.05$ indicates that our method is significantly better than the baseline method.

Method	CAMELYON16			BgIM		
	ACC	AUC	F1-Score	ACC	AUC	F1-Score
MIL-RNN	0.0086	0.0194	0.0006	$5.25e-5$	0.0001	$1.91e-5$
ABMIL	0.0011	0.0060	0.0017	$8.51e-7$	$7.48e-5$	$1.95e-7$
CLAM-MB	0.0318	0.0089	0.0193	0.0008	0.0008	$3.09e-5$
DSMIL	0.0086	0.0059	0.0015	$4.92e-5$	0.0001	$4.63e-5$
TransMIL	0.1091	0.1874	0.0451	0.0010	0.0003	0.0005
ILRA-MIL	0.1893	0.0581	0.1191	0.0002	0.0012	0.0002

- **Extra extractor.** Based on your suggestion, we report the effect of self-supervised training [WYZ+22] of swin-tiny as a feature extractor as follows:

Ablation item	CAMELYON16			BgIM		
	ACC	AUC	F1-score	ACC	AUC	F1-score
w/ ImageNet weight	$85.12_{1.58}$	$87.02_{3.87}$	$84.13_{1.22}$	$80.00_{1.90}$	$92.49_{0.97}$	$70.41_{1.05}$
w/ ctranspath weight	$85.73_{1.69}$	$88.45_{1.61}$	$84.84_{1.64}$	$77.62_{4.15}$	$92.70_{0.66}$	$70.16_{5.87}$
NcIEMIL	$\mathbf{86.05_{1.55}}$	$\mathbf{89.68_{2.10}}$	$\mathbf{85.26_{1.54}}$	$\mathbf{85.23_{0.95}}$	$\mathbf{95.87_{0.60}}$	$\mathbf{81.20_{0.94}}$

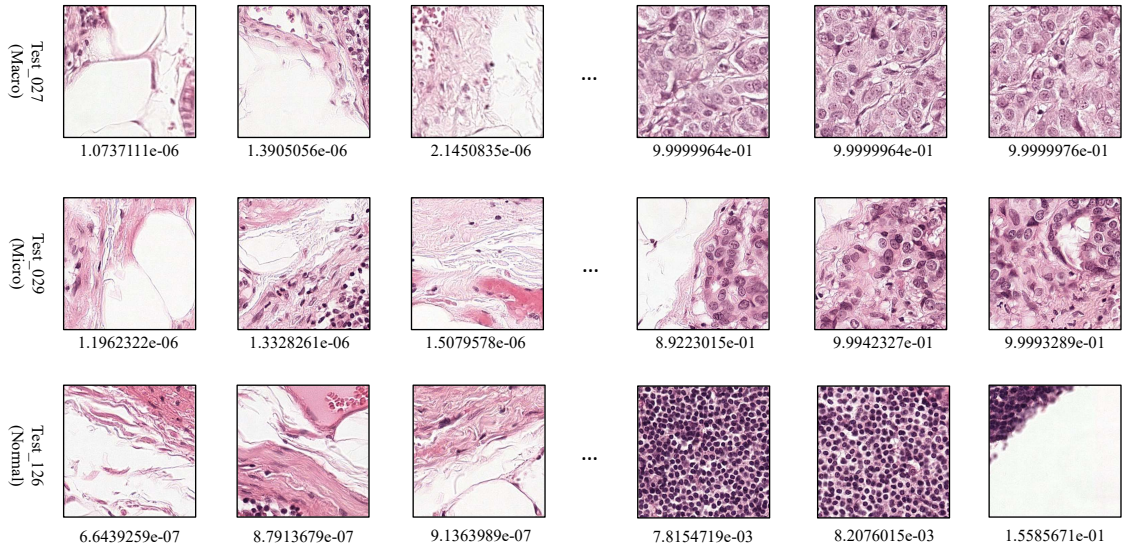

Figure 1: Visualization on Camelyon16

- **Time-consumption of pre-training.** We completely agree with your perspective. Indeed, the pre-training process of the feature extractor does require some time, falling somewhere between directly loading ImageNet weights and engaging in self-supervised learning. The feature extractor typically converges within 80 to 120 epochs. However, there are ways to address this drawback. For example, one approach could involve randomly sampling within the negative bags while keeping the positive bags untouched. Additionally, performing distributed computation could also help alleviate the time consumption.

- **Visualization.** According to your suggestion, we present the top-3 instances and the negative top-3 instances from the 4-grades slides of BgIM and normal, micro-, and macro-metastasis slides of CAMELYON16, respectively. This section will be included in the appendix.

Thanks again for your comments, and we will adjust the wording of section 2.2, although there is little relevant research on instance filtering. All additional experimental results will be added to the paper.

References

[CHG+19] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[WYZ+22] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022.

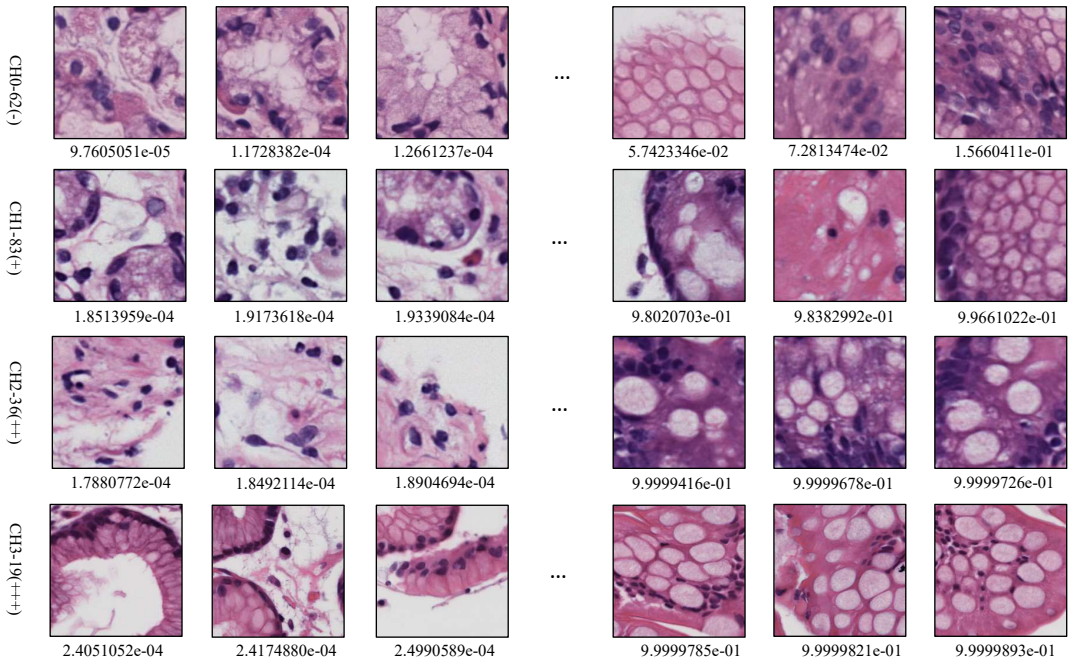

Figure 2: Visualization on BgIM