TASK-SPECIFIC FINE-TUNING VIA VARIATIONAL INFORMATION BOTTLENECK FOR WEAKLY-SUPERVISED PATHOLOGY WHOLE SLIDE IMAGE CLASSIFICATION

Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui,
Wenwei Kuang, Sunyi Zheng, Lin Yang
Westlake University, Zhejiang University.
lihonglin@westlake.edu.cn

JUNE 18-22. 2023 CVPR VANCOUVER, CANADA

Introduction

To alleviate the dilemma of computation cost and performance, we propose an efficient Whole Slide Image (WSI) fine-tuning framework motivated by the Information Bottleneck theory. The theory enables the framework to find the minimal sufficient statistics of WSI, thus supporting us to fine-tune the backbone into a task-specific representation only depending on WSI-level weak labels. The WSI Multi-instance Learning (WSI-MIL) problem is further analyzed to theoretically deduce our fine-tuning method. Our framework is evaluated on five pathology WSI datasets on various WSI heads. The experimental results of our fine-tuned representations show significant improvements in both accuracy and generalization compared with previous works.

Method

Variational Information Bottleneck

The Information Bottleneck (IB) can work as an information compression role to intervene in DNN's training [1]. The objective function of IB to be maximized is given in [7] as,

$$R_{IB} = I(Z,Y) - \beta I(Z,X), \tag{1}$$

where $I(\cdot,\cdot)$ indicates the Mutual Information (MI) and β is a Lagrange multiplier controlling the trade-off between the information that the representation variable Z shares with the label Y and its shares with input X. Since the computation of MI is intractable during the training of the neural networks, to maximize IB objective can be transferred to minimize a variational bound of Eq.(1) derived in [1] follows:

$$J_{IB} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{z\sim p_\theta(z|x_n)}[-\log q_\phi(y_n|z)] + \beta KL[p_\theta(z|x_n), r(z)], \tag{2}$$

where N denotes the number of samples, $q_\phi(y|z)$ is a parametric approximation to the likelihood $p(y|z)$, $r(z)$ is the prior probability of z to variational approximate the marginal $p(z)$, and $p_\theta(z|x)$ is the parametric posterior distribution over z.

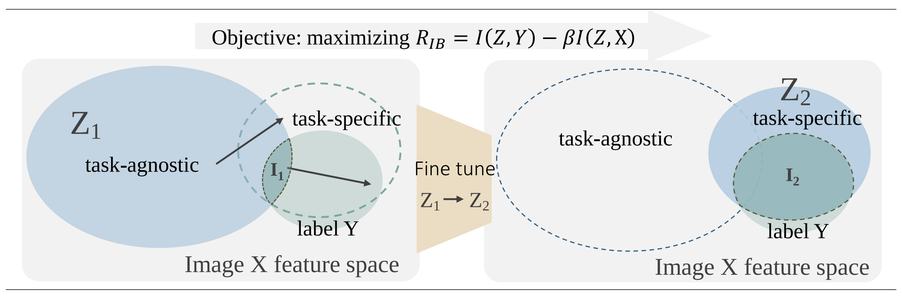

Figure 1: An illustration of the tuning scheme.

Learn MIL Sparsity via Variational Bound

The above filtering process can be implemented by optimizing the second term of in Eq.(1) which controls the compression. For the setting of our long instance sequenced MIL, we reduce $I(X,Z)$ into a degree so that the gradients can be back-propagated to the backbone encoder, which needs us to convert a WSI of bag size over 10k into 1k for the sake of sparsity. Considering MIL for tumor v.s. normal binary classification without loss of generality and the latent label y_i of each instance x_i, we argue that it is sufficient enough to make the WSI level prediction if one tumor area is detected. With the above understanding, we propose to learn compressed components similar to [5] by defining a IB module as:

$$z = m \odot x, \tag{3}$$

where m is a Bernoulli(π) distributed binary mask and in this way $KL[p_\theta(z|x), r(z)]$ in Eq.(2) can be decomposed as,

$$KL[p_\theta(m_i|x), r(m_i)] + \pi H(X), \tag{4}$$

where $H(X)$ is the entropy of X, which can be omitted during the minimization due to its constant value.

The Bernoulli(π) distribution for m fits the definition of MIL empirically: we can treat m as a latent weak prediction $\hat{y}$ describing whether the patch contains tumor or not, denoting $P_{set} = \{p(m_1|x_1), ..., p(m_N|x_N)\}$, then during inference $\hat{Y}$ can be derived as:

$$\hat{Y} = \max\{P_{set}\} = \max\{P_{subset}\}, \tag{5}$$

where $P_{subset} \in P_{set}$, generated by select top-K elements in P_{set}. The patch classifier trained with only slide-level supervision shows low accuracy [3], so we only use it to generate mask for sparse sub-bag and still utilize attention-based MIL on the sub-bag for decision making.

Fine-tuning

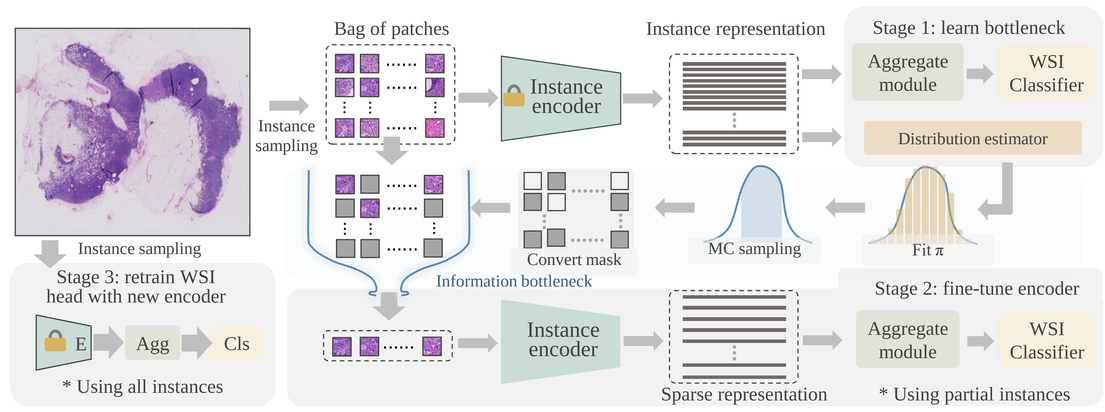

Figure 2: Workflow of WSI-MIL task-specific fine-tuning. 1) Initialize the backbone with pretrained parameters and set frozen, then learn the IB module to generate instance masks. 2) fix the mask to distill a sparse bag, then fine-tune the WSI head and patch the backbone end-2-end. 3) utilize all fine-tuned instance features within a bag and train the WSI-MIL classifier head.

Resutls

Slide-level Classification

Table 1: Slide-Level Classification by using the IN-1K pre-trained backbone or the proposed fine-tuned (FT) in three datasets. Different MIL architectures are compared to select the top 3 SOTA methods to validate the transfer learning performance using the IN-1K pre-trained backbone or the FT.

Method	Camelyon-16		TCGA-BRCA		LBP-CECA	
	F1	AUC	F1	AUC	F1	AUC
Full Supervision	0.967±0.005	0.992±0.003	-	-	0.741±0.006	0.942±0.002
AB-MIL [2]	0.828±0.013	0.851±0.025	0.771±0.040	0.869±0.037	0.525±0.017	0.845±0.002
DS-MIL[3]	0.857±0.023	0.892±0.012	0.775±0.046	0.875±0.041	-	-
CLAM-SB [4]	0.839±0.018	0.875±0.028	0.797±0.046	0.879±0.019	0.587±0.014	0.860±0.005
TransMIL [6]	0.846±0.013	0.883±0.009	0.806±0.046	0.889±0.036	0.533±0.006	0.850±0.007
DTFD-MIL [8]	0.882±0.008	0.932±0.016	0.816±0.045	0.895±0.042	0.569±0.026	0.847±0.003
FT+ CLAM-SB	0.911±0.017	0.956±0.013	0.845±0.032	0.935±0.027	0.718±0.010	0.907±0.005
FT+ TransMIL	**0.923**±0.012	**0.967**±0.003	0.848±0.044	0.945±0.020	0.720±0.024	0.918±0.004
FT+ DTFD-MIL	0.921±0.007	0.962±0.006	**0.849**±0.027	**0.951**±0.016	**0.723**±0.008	**0.922**±0.005

In addition, SSL can be combined with the proposed framework for further improvement. Compared with fully supervised learning, our methods can achieve competitive accuracy by utilizing extremely weak WSI labels. Furthermore, our training scheme can introduce versatile training-time augmentations for better generalization on datasets with domain shift, which is an inevitable challenge for previous work. The experimental results reflect the advances of our method in both accuracy and generalization.

References

[1] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[2] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

[3] B. Li, Y. Li, and K. W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[4] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[5] B. Paranjape, M. Joshi, J. Thickstun, H. Hajishirzi, and L. Zettlemoyer. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online, Nov. 2020. Association for Computational Linguistics.

[6] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and y. zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

[7] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[8] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *ArXiv*, abs/2203.12081, 2022.