Collaborative Learning with Curriculum Loss for Accurate and Interpretable Weakly Supervised Whole-Slide Image Classification Collaborative Learning with Curriculum Loss for Whole-Slide Image Classification

Rui Yang* YANG19981125@126.COM
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Pei Liu YUUKILP@163.COM
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Luping Ji JILUPING@UESTC.EDU.CN
School of Computer Science and Engineering, University of Electronic Science and Technology of China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Weakly-Supervised Learning (WSL) has been increasingly concerned in Whole-Slide Image (WSI) classification, meanwhile, an open question arises: could WSL-based models provide us with an accurate interpretation of their decisions? Although many research works have made exciting progress via building an Auxiliary Instance Branch (AIB) on a bag-level network, there are still two typical problems to be confronted with in training WSL-based AIB: i) an overwhelming influence of negative instances and ii) the inconsistent learning between bag-level network and AIB. To address them, this paper proposes collaborative learning with curriculum loss. This scheme, on one hand, provides a curriculum loss for optimizing AIB, to alleviate the first problem. Considering the knowledge reliability in WSL, this loss generalizes an original quality focal loss to WSL scenarios by curriculum instances. On the other hand, to overcome the second problem, this scheme trains a bag-level network under the supervision of AIB by a reversed curriculum loss, making both learn collaboratively. Comparative experiments prove that our scheme could often surpass existing ones in both accuracy and interpretability. Moreover, it is found that the knowledge reliability-inspired curriculum instance is a critical factor in bringing comprehensive improvements.

Keywords: Computational Pathology, Weakly-Supervised Learning, Multiple Instance Learning, Model Interpretability, Curriculum Loss

1. Introduction

The classification of whole-slide images (WSI) is challenging due to difficulties in modeling gigapixel images and training efficient deep learning (DL) models. A weakly-supervised learning algorithm, multiple instance learning (MIL), is proposed for WSI classification using weak labels and treating a single image as a bag of instances.

Embedding-level MIL (Ilse et al., 2018), exploiting mutual-instance relations to enhance bag-level representation, has achieved success by incorporating some advanced DL architectures or techniques like GCN (Liu et al., 2023), Transformer (Zheng et al., 2023), feature pyramids (Li et al., 2021) and self-supervised pre-training. However, they often lack interpretability due to the inability to explicitly infer instance probabilities, hindering their usability in clinical settings.

Some other MIL methods tackle this problem by building an auxiliary instance branch (AIB) upon embedding-level MIL (see Figure 1 (a)). ABMIL (Ilse et al., 2018) adopts an attention-based

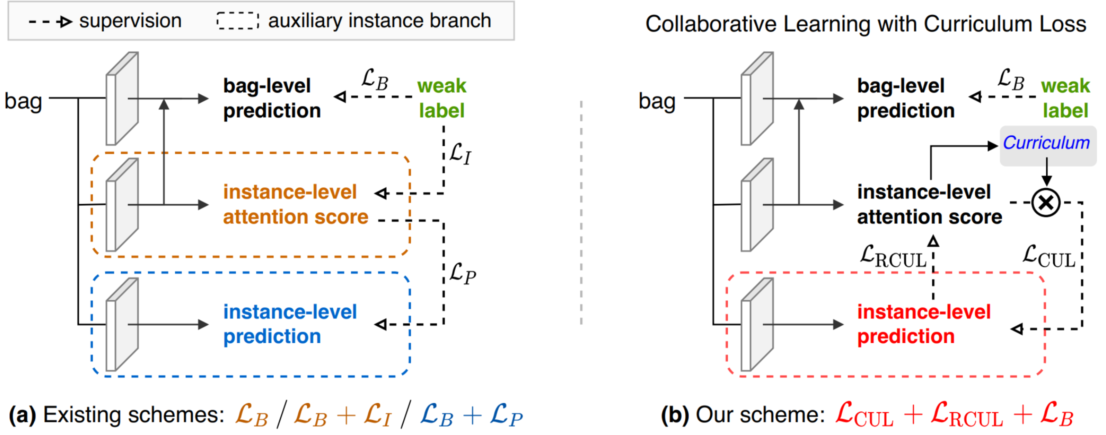
supervision
auxiliary instance branch
bag
bag-level prediction
$\mathcal{L}_B$
weak label
$\mathcal{L}_I$
instance-level attention score
$\mathcal{L}_P$
instance-level prediction
(a) Existing schemes: $\mathcal{L}_B$ / $\mathcal{L}_B + \mathcal{L}_I$ / $\mathcal{L}_B + \mathcal{L}_P$
Collaborative Learning with Curriculum Loss
bag
bag-level prediction
$\mathcal{L}_B$
weak label
Curriculum
instance-level attention score
$\mathcal{L}_{\mathrm{RCUL}}$
$\mathcal{L}_{\mathrm{CUL}}$
instance-level prediction
(b) Our scheme: $\mathcal{L}_{\mathrm{CUL}} + \mathcal{L}_{\mathrm{RCUL}} + \mathcal{L}_B$

Figure 1: Comparison of existing MIL methods with AIB and ours: (a) the typical learning scheme in existing works [1][2][6][7]; (b) our curriculum loss based collaborative learning scheme. The color of one dashed box indicates one specific scheme of building AIB.

AIB to calculate instance-level attention scores updated by a single bag-level BCE (binary cross entropy) loss. Based on this, ChiMIL (Chikontwe et al.) and Loss-Attn additionally utilize an instance-level BCE loss and an attention-weighted BCE loss to optimize AIB, respectively. However, both of them rely on the noisy instance pseudo-labels directly derived from bag labels. To mitigate this, WENO (Qu et al., 2022) extends a new AIB using attention scores as pseudo-labels.

Despite the effectiveness of AIB, two obvious problems exist in practice. (1) Overwhelming negative instances: Overproportioned negative instances tend to dominate in AIB training. It is a typical class imbalance problem that could impair network training (Li et al., 2020). (2) Inconsistent interpretability between bag-level network and AIB: AIB often has better performance in interpretability, while the bag-level network, exactly for teaching AIB, often falls behind AIB obviously. This could make AIB degenerate into a sub-optimal one.

To address the two problems analyzed above, this paper proposes a collaborative learning scheme with curriculum loss, as shown in Figure 1(b). On one hand, to tackle the problem of class imbalance in WSL scenarios, a curriculum loss is proposed for teaching balanced and reliable knowledge. On the other hand, this curriculum loss is simultaneously leveraged to conversely optimize the bag-level network, i.e., it supervises the attention output of the bag-level network, enabling the collaborative learning between the bag-level network and AIB.

Our main contributions are summarized as follows. (1) This paper introduces a collaborative learning scheme with curriculum loss for accurate and interpretable weakly-supervised WSI classification. It could not only enable AIB to be taught with balanced and reliable knowledge but also make the bag-level network and AIB learn collaboratively. (2) Through experiments, this paper demonstrates that our knowledge reliability-inspired curriculum loss could often bring comprehensive performance improvements in both WSI classification and tumor localization. And the bag-level MIL network could further benefit from our collaborative learning scheme.

2. Methodology

2.1. Preliminaries

Assume a bag $X = \{x_1, \ldots, x_K\}$ of K instances where $X_k \in R^d$ for $k = 1, \ldots, K$. The bag-level label of X is denoted as $Y \in \{0, 1\}$. Let ϕ represents a bag-level network and f be its attention branch. Instance attention scores are $A = f(X)$, where A_k is the attention score of the k^{th} instance. And bag-level prediction is $Y_b = \phi(X, A)$. To interpret $\hat{Y}$, f is adopted as AIB in ABMIL (Ilse et al., 2018), ChiMIL (Chikontwe et al.), and Loss-Attn (Shi et al., 2020). ABMIL adopts a single bag-level BCE loss as follows,

$$\mathcal{L}_B = -((1 - Y)\log(1 - \hat{Y}) + Y\log(\hat{Y})) \tag{1}$$

Casting Y as the pseudo-label of all instances, ChiMIL additionally uses an instance-level BCE loss (see Figure 1(a)),

$$\mathcal{L}_I = -\sum_{k=1}^{K}\left((1 - Y)\log\left(1 - \mathrm{S_k}\right) + Y\log\left(S_k\right)\right) \tag{2}$$

where $S_k = \mathrm{Norm}\left(A_k\right) \in [0, 1]$ is the normalized attention score of the k^{th} instance. In WENO, a new AIB g is built and outputs the instance prediction $\hat{y}_k = g\left(x_k\right)$. It denotes the loss function as,

$$\mathcal{L}_P = -\sum_{k=1}^{K}\left((1 - S_k)\log\left(1 - \hat{y}_k\right) + S_k\log\left(\hat{y}_k\right)\right) \tag{3}$$

where S_k rather than Y, is taken as the instance pseudo-label for training AIB.

Original QFL (Li et al., 2020) is proposed to alleviate the problem of class imbalance in training AIB. It can be written as follows:

$$\mathcal{L}_{QFL} = -\sum_{k=1}^{K}\left|S_k - \hat{y}_k\right|^{\beta}\left((1 - S_k)\log\left(1 - \hat{y}_k\right) + S_k\log\left(\hat{y}_k\right)\right) \tag{4}$$

where $\beta \geq 0$ is a hyper-parameter. Compared to $\mathcal{L}_P$, $\mathcal{L}_{QFL}$ has a new term $|S_k - \widehat{y_k}|^{\beta}$ that acts as a modulating factor. This term imposes more attention on those instances of which prediction and pseudo-labels deviate significantly. However, this new term is not always reliable as it relies heavily on the assumption that S_k is exactly equivalent to ground truth. Next, we show how to address this problem so as to make QFL well adapt to WSL.

2.2. Collaborative Learning with Curriculum Loss

2.2.1. CURRICULUM LOSS

As shown in Figure 2 (a), considering the reliability of instance pseudo-labels, we propose a curriculum-based for training AIB, inspired by curriculum learning. We argue that instance attention scores are often uncertain in prediction and could not be directly taken as reliable knowledge to teach an AIB, which is confirmed through our ablation studies.

At first, we cast instance as curriculum and employ a scoring function $\sigma(\cdot)$ to assess the curriculum hardness on S_k, i.e., $h_k = \sigma\left(S_k\right) \in [0, 1]$. Then, we adopt h_k (curriculum hardness score) to

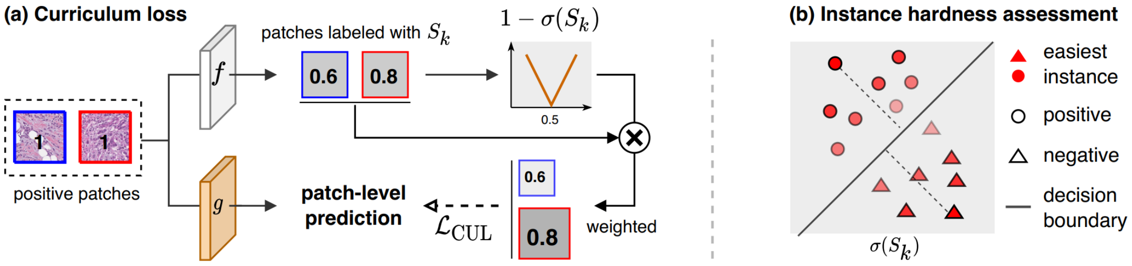

Figure 2: (a) Illustration of the curriculum loss and (b) Illustration of function $\sigma(\cdot)$.

adjust the instance weight in training AIB. Thereby, the curriculum loss (CUL), a generalized QFL for weakly-supervised WSI classification, is written as:

$$\mathcal{L}_{\text{CUL}} = -\sum_{k=1}^{K} \left(1-\sigma\left(S_k\right)\right) \left|S_k - \hat{y}_k\right|^{\beta} \left(\left(1-S_k\right)\log\left(1-\hat{y}_k\right) + S_k \log\left(\hat{y}_k\right)\right) \tag{5}$$

where $1-\sigma\left(S_k\right)$ is a new weighting factor employed to weaken the effect of QFL when it is a hard instance (i.e., uncertain knowledge). This new factor could encourage AIB to preferentially put focus on easy curriculums and accordingly be taught with more reliable knowledge.

We design a simple and intuitive form for this new factor, $1-\sigma\left(S_k\right) = [2S_k - 1\,|$. Namely, instance hardness is assessed by $\sigma\left(S_k\right) = 1 - [2S_k - 1\,|$. This $\sigma(\cdot)$ implies that instance hardness directly relies on the distance from an instance to a decision boundary. And the closer a distance is, the harder an instance is, as illustrated in Figure 2(b). Note that this form is based on an assumption that there is a decision boundary separating positive and negative instances in the hypothesis space of f. Despite the simplicity and constraint of $\sigma(\cdot)$, we observe its favorable performance in experiments when applying it to QFL.

2.2.2. Collaborative Learning

To mitigate the problem of inconsistent learning, we further utilize a reversed CUL to optimize the attention output of bag-level network and make bag-level network and AIB learn collaboratively. Specifically, we employ AIB (better in instance interpretability) to supervise attention scores. This reversed CUL (RCUL) can be presented by

$$\mathcal{L}_{RCUL} = -\sum_{k=1}^{K} \left(1-\sigma\left(\hat{y}_k\right)\right) \left|\hat{y}_k - S_k\right|^{\beta} \left(\left(1-\hat{y}_k\right)\log\left(1-S_k\right) + \hat{y}_k \log\left(S_k\right)\right) \tag{6}$$

In our scheme, the total loss for weakly-supervised WSI classification is $\mathcal{L}_{CUL} = \mathcal{L}_B + \mathcal{L}_{CUL} + \mathcal{L}_{RCUL}$, as shown in Figure 1(b). It is optimized in a bi-directional knowledge distillation manner (Qu et al., 2022; Wang et al., 2021). Specifically, in each training epoch, ϕ and f are firstly optimized by minimizing $\mathcal{L}_B + \mathcal{L}_{CUL}$ before g, and then g is optimized by minimizing $\mathcal{L}_{CUL}$. Moreover, in evaluation, bag-level and instance-level predictions are derived from ϕ and g, respectively.

3. Experiments and Results

3.1. Experimental settings

Two histopathology datasets, Camelyon16 (Ehteshami Bejnordi et al., 2017) and DigestPath2019 (Da et al., 2022) are used for experiments. In the former two sub-datasets with different magnifications are chosen, while the latter has only one resolution.

For model evaluation, AUC (area under the curve) is used to measure overall performance in both patch-level and slide-level classification. The FROC (free response operating characteristic), a widely-used metric in tumor detection, is reported to quantify the interpretability of WSL models.

We compare our method to the embedding-level MIL ones with AIB (Chikontwe et al.; Ilse et al., 2018; Qu et al., 2022; Shi et al., 2020). Following WENO [6], we use a same g and adopt the ABMIL network to implement ϕ and f. We reproduce all these methods on their released codes. β is set to 1 by default. All experiments run on a machine with 2 GeForce-RTX3080Ti (12G) GPUs.

3.2. Overall performance

From the comparative results shown in Table 1, we can find that our method could often perform better than the compared ones in both tumor localization and slide-level classification in both datasets. Specifically, on Camelyon16, our method could surpass most existing ones, except for Loss-Attn, in patch-level classification. The improvement on DigestPath2019 is more obvious with FROC increased by 4.55%-11.06%, patch-level AUC increased by 3.15%-12.76, and slide-level AUC increased by 0.3%-2.52%. These results suggest that our method enjoys both accuracy and interpretability in WSL-based WSI classification.

Table 1: Weakly-supervised tumor localization and classification performance on two datasets

Method	Camelyon16(5×)			Camelyon16(20×)			DigestPath2019		
	FROC	Patch AUC	Slide AUC	FROC	Patch AUC	Slide AUC	FROC	Patch AUC	Slide AUC
ABMIL	0.6970	0.9384	0.8419	0.4425	0.7535	0.8803	0.2471	0.7720	0.9405
ChiMIL	0.7627	0.9272	0.7301	0.5363	0.8966	0.7278	0.2745	0.8683	0.9524
Loss-Attn	0.7805	**0.9722**	0.7919	0.5539	**0.9551**	0.8226	0.2094	0.8344	0.9302
WENO	0.7504	0.8858	0.8059	0.4992	0.9045	0.8255	0.2482	0.9305	0.9332
Ours	**0.7863**	0.9518	**0.8842**	**0.5590**	0.9173	**0.9152**	**0.3200**	**0.9620**	**0.9554**

3.3. Ablation study

Ablation study We conduct ablation studies to further understand the effect of $\mathcal{L}_{CUL}$ and $\mathcal{L}_{RCUL}$. A weighted BCE loss adopted by the original WENO is taken as the baseline for comparisons. We have some notable findings from the results shown in Table 2. (1) Original QFL performs worse than the BCE loss in the classification of WSIs at 20×. This result backs up our aforementioned argument, i.e., original QFL could lose its functionality when the assumption that the instance pseudo label is exactly equivalent to ground truth is not reliable. (2) The curriculum instance, as the core of L_CUL to generalize QFL, indeed is a critical factor in achieving favorable performances, since we observe its comprehensive improvements over QFL, especially on the WSIs at 20×. This fact suggests that our curriculum loss could be more suitable for WSL, or in other words, knowledge reliability should be noticed and addressed when teaching a weakly supervised instance branch. (3) The collaborative learning, facilitated by $\mathcal{L}_{RCUL}$, could often contribute to slide-level classification but may degenerate patch-level performance (e.g., on 20×).

Table 2: Weakly-supervised tumor localization and classification performance on two datasets

scale	QFL	Curriculum	Collaborative	Localization	Classification	
				FROC	Patch-level AUC	Slide-level AUC
5×				0.7504	0.8858	0.8059
	✓			0.7636	0.9513	0.8320
	✓	✓		0.7642	**0.9549**	0.8779
	✓	✓	✓	**0.7863**	0.9518	**0.8842**
20×				0.4992	0.9045	0.8255
	✓			0.5260	0.8684	0.8003
	✓	✓		**0.5869**	**0.9395**	0.8641
	✓	✓	✓	0.5590	0.9173	**0.9152**

4. Discussion and Conclusion

Although many studies achieve success in WSL-based WSI classification, they usually lack good interpretability. This may weaken its usability in clinical practices. This paper summarizes the paradigm of interpretable WSL for WSI classification and enhances it by addressing existing problems. Meanwhile, it is found that knowledge reliability should be noticed and addressed when teaching a WSL-based AIB. Still, our work has some constraints, such as i) the limited WSI dataset for interpretability assessment and ii) more granularity for a comprehensive evaluation of model interpretability and more attempts on curriculum scoring functions (Wang et al., 2021) are anticipated.

This paper proposes a scheme of collaborative learning with curriculum loss for accurate and interpretable weakly-supervised WSI classification. It aims at mitigating the problems of overwhelming negative instances and inconsistent learning to enhance AIB. Experiments show that our scheme is effective in multiple instance learning framework, and could often achieve state-of-the-art performance in terms of both accuracy and interpretability. Moreover, our scheme could localize tumor regions more accurately than other compared methods.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities with Grant number ZYGX2022YGRH015.

References

Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. Medical Image

Computing and Computer Assisted Intervention – MICCAI 2020, pages 519–528. Springer International Publishing.

Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, and et al. Zhiqiang Hu. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80: 102485, 2022. doi: 10.1016/j.media.2022.102485.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017. doi: 10.1001/jama.2017.14585.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2021. doi: 10.1109/CVPR46437.2021.01409.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, 2020.

Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Graphlsurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. *Computer Methods and Programs in Biomedicine*, 231:107433, 2023. doi: 10.1016/j.cmpb.2023.107433.

Linhao Qu, Xiaoyuan Luo, Manning Wang, and Zhijian Song. Bi-directional weakly supervised knowledge distillation for whole slide image classification, 2022.

Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. In *AAAI Conference on Artificial Intelligence*, 2020.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning, 2021.

Yushan Zheng, Jun Li, Jun Shi, Fengying Xie, Jianguo Huai, Ming Cao, and Zhiguo Jiang. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. *IEEE Transactions on Medical Imaging*, 42(9):2726–2739, 2023. doi: 10.1109/TMI.2023.3264781.