MULTIPLE INSTANCE LEARNING WITH TASK-SPECIFIC MULTI-LEVEL FEATURES FOR WEAKLY ANNOTATED HISTOPATHOLOGICAL IMAGE CLASSIFICATION

Yuanpin Zhou[⋆] *Yao Lu*[⋆†]

[⋆] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, P.R. China
[†] Perception Vision Medical Technology Company Ltd., Guangzhou 510275, P.R. China

ABSTRACT

Pathological examination is regarded as the gold standard for cancer diagnosis in clinical medicine. Gigapixel histopathological images play an important role in developing automatic pathology diagnosis systems. Because of the expensive cost of pixel-level annotation, deep multiple instance learning (MIL) has become the main approach to classify histopathological images with only slide-level annotation. However, three major challenges including lack of data efficiency because MIL approaches rely on task-agnostic feature extractor, overfitting challenges caused by high data imbalance between tumor and normal tissues, and the similarity between tumor and normal patches, are to be tackled. We proposed a three-stage deep MIL approach to address these challenges. The first stage generated pseudo instance-level labels by utilizing max-max ranking loss. The second stage refined the feature extractor into a task-specific feature extractor with pseudo labels. Focal loss was used to tackle the data imbalance challenge. Partial decoder component was adopted to better distinguish tumor tissues. We re-trained the MIL network with task-specific multi-level features in the third stage. Our model was able to predict both instance-level and bag-level labels in the inference stage. Experiment results showed our model outperformed baseline models in both accuracy and AUC metrics.

Index Terms— Multiple instance learning, weak supervision, deep learning, histopathology, whole slide images

1. INTRODUCTION

Histopathological image has been regarded as the gold standard for diagnosing cancer in clinical medicine for several years. A specialized scanner is used to convert entire histopathology on a glass slide into a digital whole slide image (WSI). Extremely high resolution WSIs which typically contain trillions of pixels are generated by this technique. Along with the rich information for cancer diagnosis benefited from the extremely high resolution, WSIs also bring challenges to developing an automatic diagnosis system of histopathology.

One of the biggest challenges in histopathological image analysis with WSIs is the expensive cost to obtain pixel-level labels by expert pathologists due to the gigapixel resolution. Most WSIs are weakly annotated, which means only slide-level labels (normal or tumor) are available. To address this challenge, one of the main approaches is to consider the weakly annotated WSI classification problem as a multiple instance learning (MIL)[1] problem, where each WSI is considered as a bag that contains many patches named instances. MIL is a typical form of weakly-supervised learning algorithms where only bag-level labels are available for training. Under the standard MIL assumption [2], all benign bags (in our case are normal WSIs) contain only benign instances, and that malignant bags (tumor WSIs) contain at least one malignant instance. Recently, benefited from using deep neural networks, deep MIL based approaches[3, 4, 5] brought promising results to weakly annotated histopathological image classification.

However, three major challenges still exist in developing weakly annotated histopathological image classification model under the deep MIL framework. Firstly, current deep MIL model for histopathology rely on extracting deep features from a task-agnostic pre-trained deep neural networks as end-to-end training a feature extractor is prohibitively expensive for large bags[6]. Secondly, deep MIL models can suffer from overfitting because of the low witness rate[2] characteristic in histopathology, which means most patches are normal tissues even in a tumor WSI. Thirdly, as shown in Fig. 1, the boundary between normal tissues and tumor tissues is indeterminate as tumor tissues "seamlessly" embedded in their surroundings. It requires a more sophisticated design of deep neural network architecture.

To address these challenges, we propose a multiple instance learning network with task-specific multi-level features for WSI classification. A three-stage training procedure that allowed us to refine the feature extractor was introduced to tackle the first challenge of task-agnostic feature extraction. The first stage was to generate pseudo instance-level labels by deep MIL framework. The second stage was to refine the feature extractor into a task-specific feature extractor with pseudo instance-level labels. We used focal loss[7] for task-specific feature extractor training to tackle the low

witness rate challenge. Moreover, we adopted the partial decoder component (PDC)[8], which had been proven to be effective in camouflaged object detection (COD)[9], in the task-specific feature extractor architecture to tackle the third challenge. In the third stage, we re-trained the MIL network with task-specific multi-level features generated by previous stages. In the inference stage, WSIs were cropped into patches and were fed into our task-specific multi-level multiple instance learning network to predict both instance-level and bag-level labels.

We evaluate our model in one of the biggest public histopathology datasets CAMELYON16[10] which contained 701 GB of histopathological images. Attention-based MIL[3] with vanilla feature extractor and multi-level concatenated feature extractor were adopted as baseline models. Experiment results showed our model outperformed baseline models by both accuracy and AUC metrics.

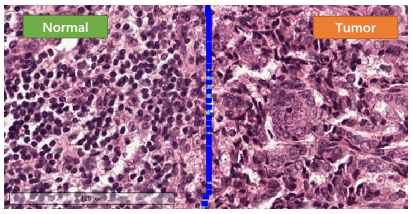

Fig. 1. An example of normal and tumor tissues in WSI

2. METHODOLOGY

The overall architecture is presented in Fig. 2. Our proposed TSML-MIL model contained three stages in training. The first stage presented in Fig. 3 was to generate pseudo instance-level labels by Deep MIL framework with max-max ranking loss (MM-RL)[11]. A WSI bags pair consists of one tumor bag and one normal bag was taken as input. Binary cross entropy loss (BCE) was utilized for bag-level label prediction respectively for tumor bag and normal bag. MM-RL was utilized for the constraint between tumor bag and normal bag. The second stage presented in Fig. 4 was to refine the feature extractor into a task-specific feature extractor with pseudo instance-level labels. Focal loss[7] was used for task-specific feature extractor training. We adopted the partial decoder component (PDC)[8], which had been proven to be effective in camouflaged object detection (COD)[9], in the task-specific feature extractor architecture. Moreover, an instance-level prediction head consists of two fully connected layers, and a softmax layer was designed to predict instance score. In the third stage, we re-trained the MIL network with task-specific multi-level features generated by previous stages. In the inference stage, WSIs were cropped into patches and were fed into our task-specific multi-level multiple instance learning network to predict both instance-level and bag-level labels.

2.1. Pseudo Label Generation

2.1.1. Attention-based MIL

Given a bag of instances $X = \{x_k\}_{k=1}^{K}$ and a feature extractor $f(\cdot)$, the feature extractor $h_k = f(x_k)$ transform instance x_k into feature embedding h_k. The purpose of the attention mechanism[3] is to learn the attention $\{\alpha_k\}_{k=1}^{K}$ within $\{h_k\}_{k=1}^{K}$. z is the linear combination of $\{h_k\}_{k=1}^{K}$. The attention mechanism could be formulated as:

$$z = \sum_{k=1}^{K} \alpha_k h_k \tag{1}$$

where:

$$\alpha_k = \frac{exp\{w^T tanh(Vh_k^T)\}}{\sum_{j=1}^{K} exp\{w^T tanh(Vh_j^T)\}} \tag{2}$$

Here $\{\alpha_k\}_{k=1}^{K}$, $w \in R^{L\times 1}$ and $V \in R^{L\times M}$ are the to-be-learnt parameters. Hyperbolic tangent $tanh(\cdot)$ introduce element-wise non-linearity to the architecture. z represents the feature of the whole bag X and is later fed into a simple fully connected neural network to predict the score of the bag.

2.1.2. Max-Max Ranking Loss

Max-max ranking loss (MM-RL) was firstly proposed for video highlight detection[11]. Fa-Ting Hong, etc. utilized a MIL framework to detect video highlights with weakly-annotation. We considered tumor tissues as video highlights and normal tissues as non-highlight segments. The following inequation can be established, where ε_p^i and ε_n^i were instance labels receptively in a positive (tumor) bag and a negative (normal) bag, $\mathcal{I}_p^i$ and $\mathcal{I}_n^i$ indicated instances and $\mathcal{B}_p$ and $\mathcal{B}_n$ indicated bags.

$$max_{\mathcal{I}_p^i \in \mathcal{B}_p} \varepsilon_p^i > max_{\mathcal{I}_n^i \in \mathcal{B}_n} \varepsilon_n^i \tag{3}$$

The above constraint between positive bags and negative bags can be formated as the max-max ranking loss (MM-RL) as:

$$\mathcal{L}_{MM}(\mathcal{B}_p, \mathcal{B}_n) = max(0, \epsilon - max_{\mathcal{I}_p^i \in \mathcal{B}_p} \varepsilon_p^i + max_{\mathcal{I}_n^i \in \mathcal{B}_n} \varepsilon_n^i) \tag{4}$$

where ϵ should be a small constant represented the gap of the maximum instance score between a positive bag and a negative bag. We set ε to 1 in our experiments.

2.2. Task-specific Multi-level Features Aggregator

2.2.1. Partial Decoder Component

As was shown in Fig. 1, tumor tissues can be difficult to distinguish from normal tissues as tumor tissues were concealed in WSIs. Recently, a new task named camouflaged object detection (COD)[9] was proposed to identify objects

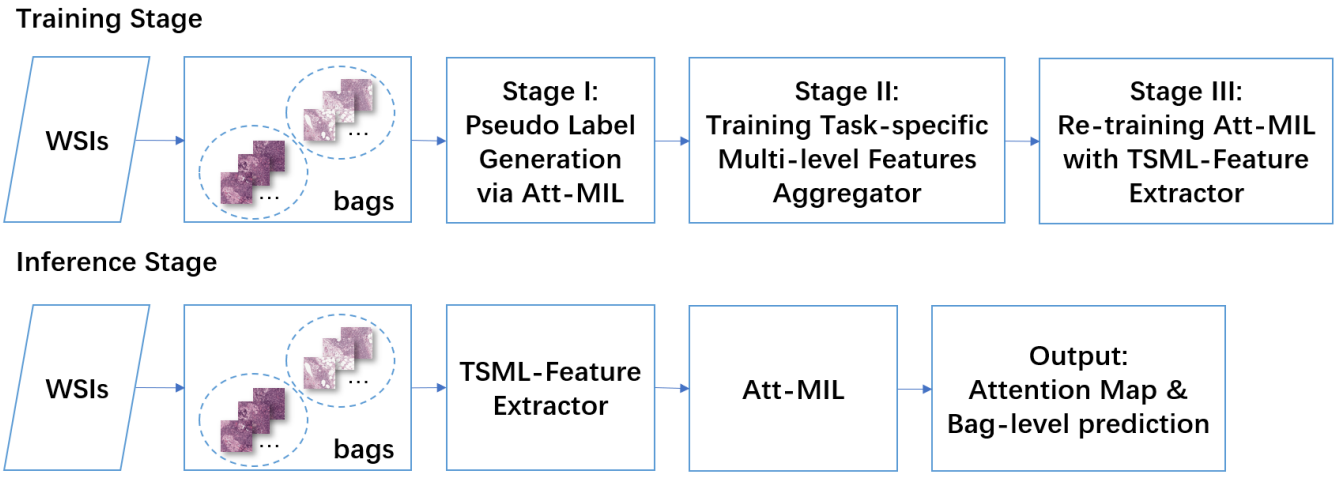

Fig. 2. Workflow of our proposed TSML-MIL framework

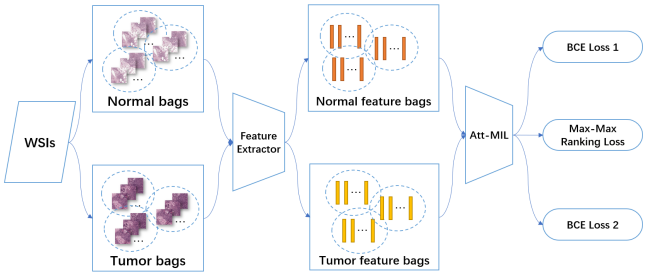

Fig. 3. Stage I: Pseudo Label Generation

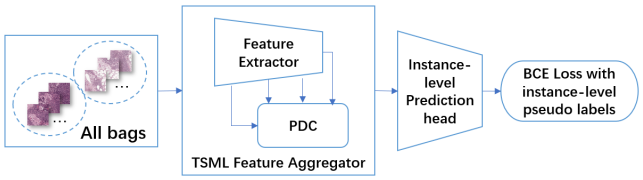

Fig. 4. Stage II: Training Task-specific Multi-level Features Aggregator

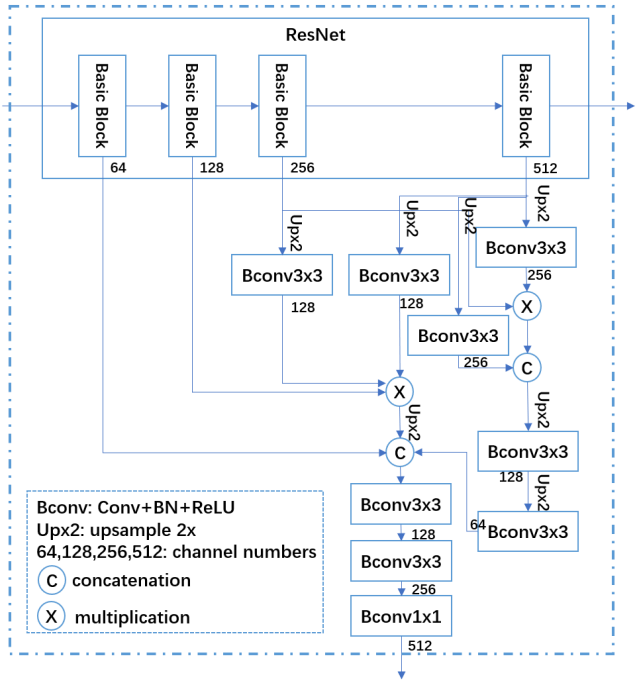

Fig. 5. Architecture of partial decoder component (PDC)

that are "seamlessly" embedded in their surroundings. As the high intrinsic similarities between the target object and the background make COD far more challenging than the traditional object detection task, addressing camouflaged object detection (COD) requires a significant amount of visual perception[12] knowledge. An extended version of partial decoder component (PDC)[8] was utilized to precisely detect the camouflaged object in the study[9].

Inspired by this, we combined the PDC module with task-agnostic pre-trained feature extractor in Stage II. The detailed architecture of our proposed PDC module was presented in Fig. 5. Since we utilized ResNet as our feature extractor, four cascaded down-sampled feature layers were adopted for PDC module. During Stage II training, the parameters of task-agnostic pre-trained feature extractor were fixed, PDC module was trained to aggregate multi-level features.

2.2.2. *Focal Loss*

Focal loss[7] was firstly proposed to tackle the challenge of extreme foreground-background class imbalance encountered during training of dense detectors. The method proposed to address class imbalance by reshaping the standard cross en-

tropy loss such that it down-weights the loss assigned to well-classified examples. Similar to foreground-background class imbalance encountered during training of dense detectors, histopathological images suffered from the same class imbalance challenge when we consider tumor patches as foreground and normal patches as background. Hence we adopted the α-balanced variant of focal loss with soft labels as our objective in Stage II.

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}[ylog(p) + (1-y)log(1-p)] \quad (5)$$

where,

$$p_t = \begin{cases} p, & if y > 0.5, \\ 1-p, & otherwise \end{cases} \quad (6)$$

Here, $\gamma \geq 0$ is a tunable focusing parameter, p denotes the instance score predicted by Stage II predictor and y denotes the instance pseudo label generated in Stage I.

3. EXPERIMENTS

3.1. Dataset

The CAMELYON16 dataset was firstly published in 2016 as a challenge of the ISBI conference for metastasis detection in breast cancer[10]. The dataset consists of 271 training WSIs and 129 testing WSIs, which yield roughly 3.2 million patches at 20× magnification and 0.25 million patches at 5× magnification with on average about 8,000 and 625 patches per bag. Tumor regions are fully annotated with pixel-level labels on each slide. We ignore the pixel-level labels and annotated slide-level labels under standard MIL assumption[2] in the dataset.

3.2. Baselines

We adopted Attention-based MIL (Att-MIL)[3] and MI-Net[13] as our backbone MIL models. For feature extraction, we utilized ResNet18[14] pre-trained on ImageNet[15] as our task-agnostic feature extractor. Two feature extraction methods were adopted in our experiments. With the vanilla method, we adopted the last average pooling layer as our extracted deep features. With the concatenated method, an additional average pooling layer was applied after each Basic Block[14], all features extracted from average pooling layers were concatenated as new features.

3.3. Experimental Analysis

The slide-level classification results were summarized in Table 1. Features were extracted on the 5x patches under the same settings. Experiment results showed that our proposed method outperformed four baseline models in multiple evaluation metrics. Experiment results also showed that the concatenated features extraction method wasn't necessarily better than the vanilla feature extraction method. The first two stages of our model trained a task-specific feature extractor. Hence the model trained in Stage3 is identical to vanilla Attention-based MIL when we take out the feature extractor. Both accuracy and AUC increased comparing TSML-MIL to Att-MIL-vanilla. Therefore, the task-specific multi-level features were better than task-agnostic features in MIL.

Moreover, our proposed model which was trained with only slide-level labels showed promising results of interpretability. A WSI example from CAMELYON16 dataset testing set was demonstrated in Fig. 6

Table 1. Experiments results

Methods	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUC*
MI-Net-vanilla	0.7132	0.6875	0.4490	0.5432	0.7018
MI-Net-concatenated	0.7209	0.7097	0.4490	0.5500	0.7173
Att-MIL-vanilla	0.7364	0.7419	0.4694	0.5750	0.7443
Att-MIL-concatenated	0.7442	0.7500	0.4898	0.5926	0.7243
TSML-MIL(ours)	**0.7829**	**0.7838**	**0.5918**	**0.6744**	**0.7912**

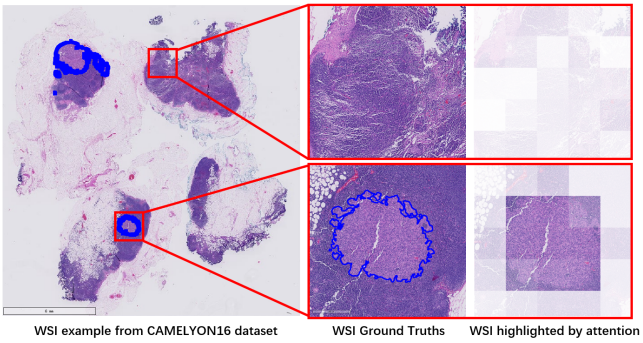

Fig. 6. Interpretability of TSML-MIL

4. CONCLUSION

In this paper, we propose a novel TSML-MIL framework for weakly annotated histopathological image classification to address three major challenges in the deep MIL framework, including relying on task-agnostic pre-trained feature extractor, highly imbalanced instance data due to low witness rate in histopathology and tumor tissues seamlessly embedded in normal tissue background. A three-stages training framework with a task-specific feature extractor embedded with partial decoder component (PDC) trained by focal loss was designed to tackle these challenges. Our model was evaluated on the 701 GB gigapixel CAMELYON16 dataset and outperformed two baseline models based on Attention-based MIL in both accuracy and AUC metrics.

5. REFERENCES

[1] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[2] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[3] Maximilian Ilse, Jakub Tomczak, and Max Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[4] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 685–694.

[5] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck, "Deep learning for identifying metastatic breast cancer," *arXiv preprint arXiv:1606.05718*, 2016.

[6] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol, "Self-supervision closes the gap between weak and strong supervision in histology," *arXiv preprint arXiv:2012.03583*, 2020.

[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[8] Zhe Wu, Li Su, and Qingming Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[9] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao, "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2777–2787.

[10] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[11] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng, "Mini-net: Multiple instance ranking network for video highlight detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 345–360.

[12] Tom Troscianko, Christopher P Benton, P George Lovell, David J Tolhurst, and Zygmunt Pizlo, "Camouflage and visual perception," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1516, pp. 449–461, 2009.

[13] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.