香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Multi-scale Domain-adversarial Multiple-instance CNN for Cancer Subtype Classification with Unannotated Histopathological Images[1]

Rongjun Tang

2021.09.08

[1] Hashimoto N, Fukushima D, Koga R, et al. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3852-3861.

Introduction

Cancer subtype classification problems:

(1) Tumor and non-tumor regions are mixed in a WSI (wide slide image, usually 40000*40000)——high cost in labeling patch.

(2) Staining conditions vary greatly depending on the specimen conditions and the hospital from which the specimen was take——system deviation.

(3) Different features of tissues are observed when the magnification of the pathological image is changed——local feature and global feature.

Introduction

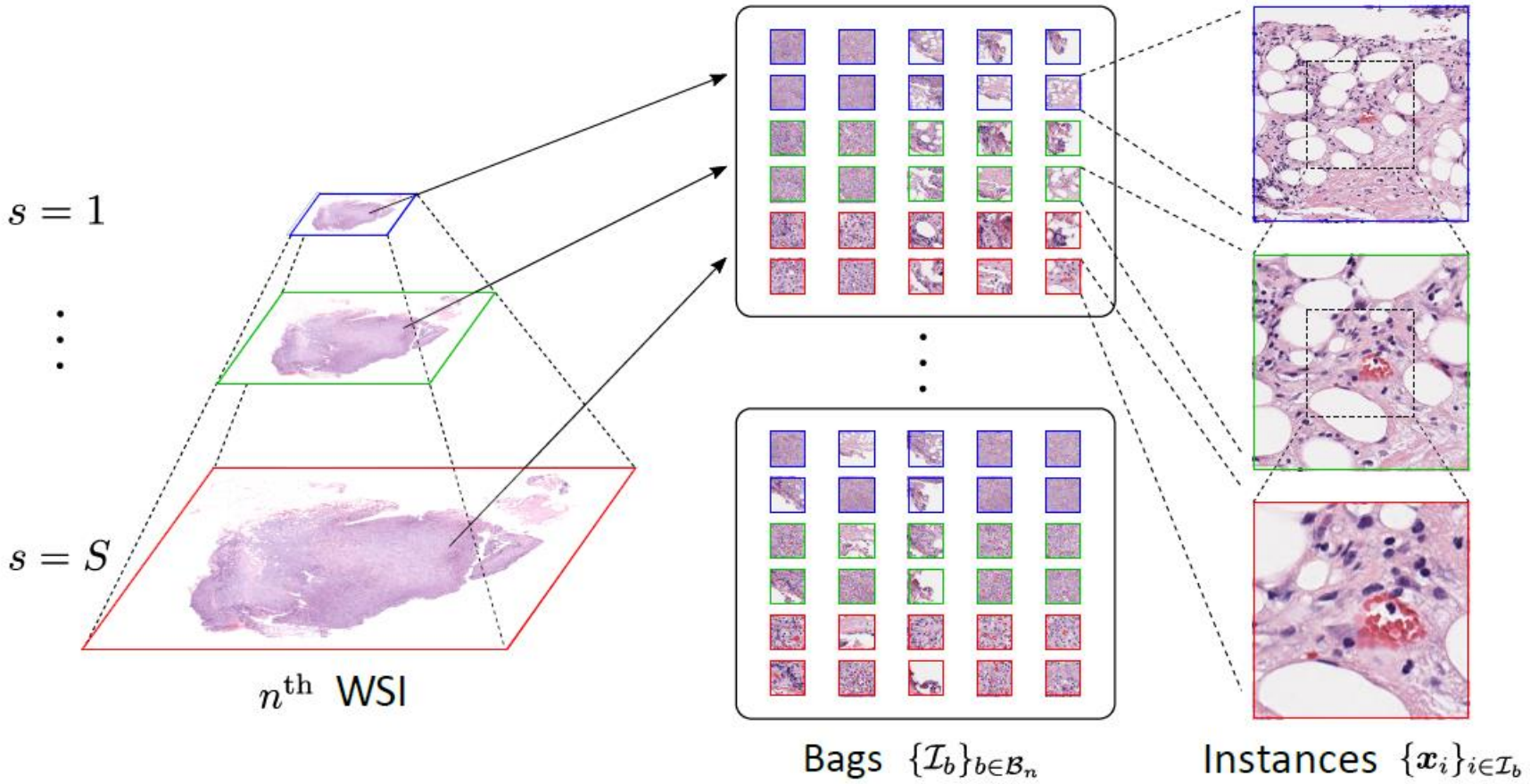
To mimic the pathologists' real practice, a CNN-based learning mechanism is proposed, which combines Multiple Instance Learning (MIL), domain adversarial (DA) normalization, and multi-scale (MS) learning techniques.

Problem Setting: Consider a training set for a binary pathological image classification problem obtained from N patients. Consider patches with 224×224 pixels. Use the idea in MIL, consider a group of patches, and assume that each group from a positive class slide contains at least a few patches having positive class-specific information, whereas each group from a negative class slide does not contain any patches having positive-class specific information. Consider patches with multiple different scales. A binary classification problem.

$s \in [S]$: index of scales.
$\mathcal{B}_n \; for \; n \in [N]$: the set of bags in n^{th} WSI and $b \in \mathcal{B}_n$ is a bag characterized by a set of patches.
$\mathcal{I}_b^{(s)}$: the patches taken from scale s and belongs to bag b.

Introduction

Illustration of notions of a WSI, bags, instances (patches), and scales:



Preliminaries: MIL

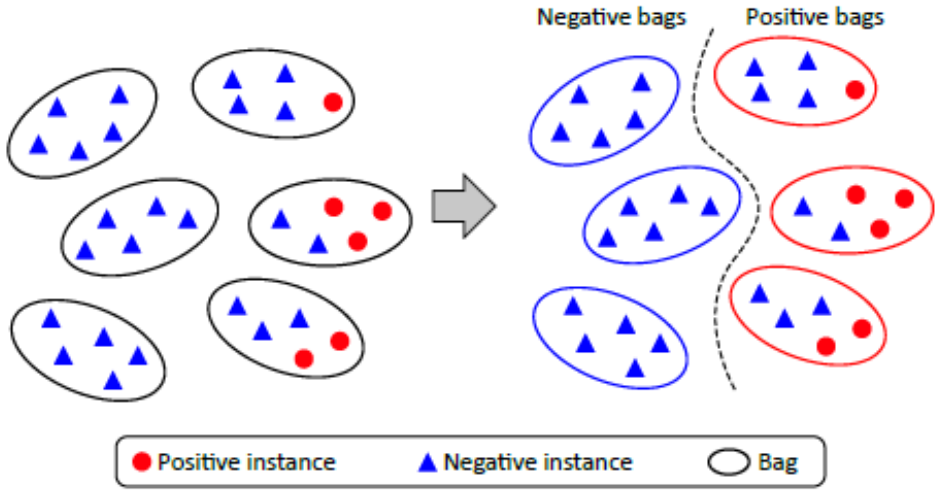


Figure 2: Explanation of MIL. Positive bags are generated from WSIs with positive subtype labels and negative bags are generated from WSIs with negative subtype labels. Only the image patches of class-specific regions, such as tumors in positive-class WSIs, are regarded as positive instances.

MIL is a type of weakly supervised learning problem, where instance labels are not observed but labels for groups of instances called bags are observed.

Preliminaries: Domain-adversarial neural network

Slide-wise differences in staining conditions, as illustrated below, highly degrade the classification accuracy. Proposed method before: color normalization and color augmentation.

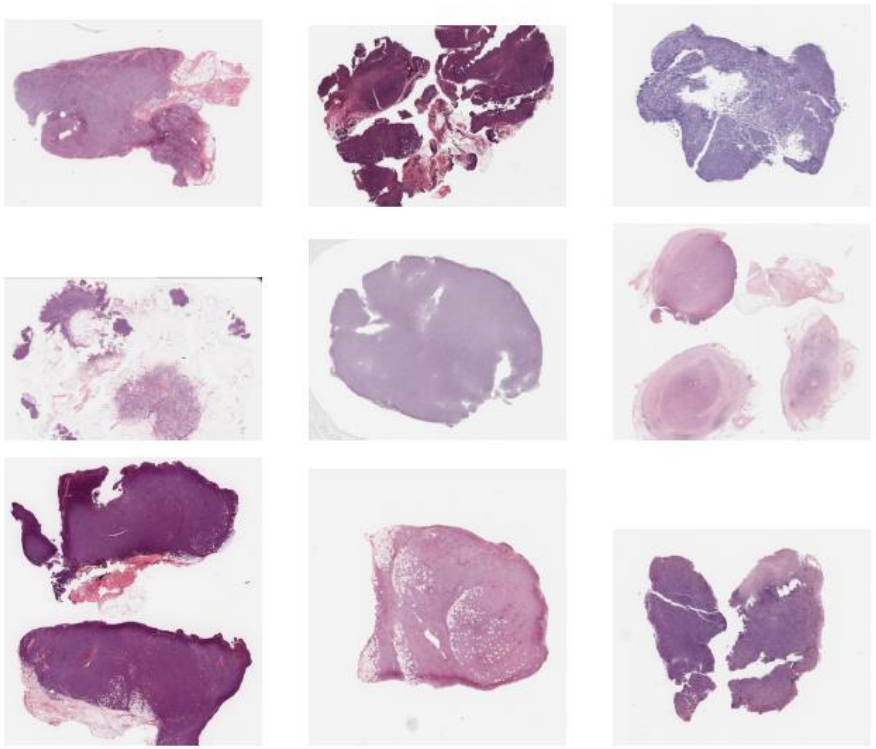


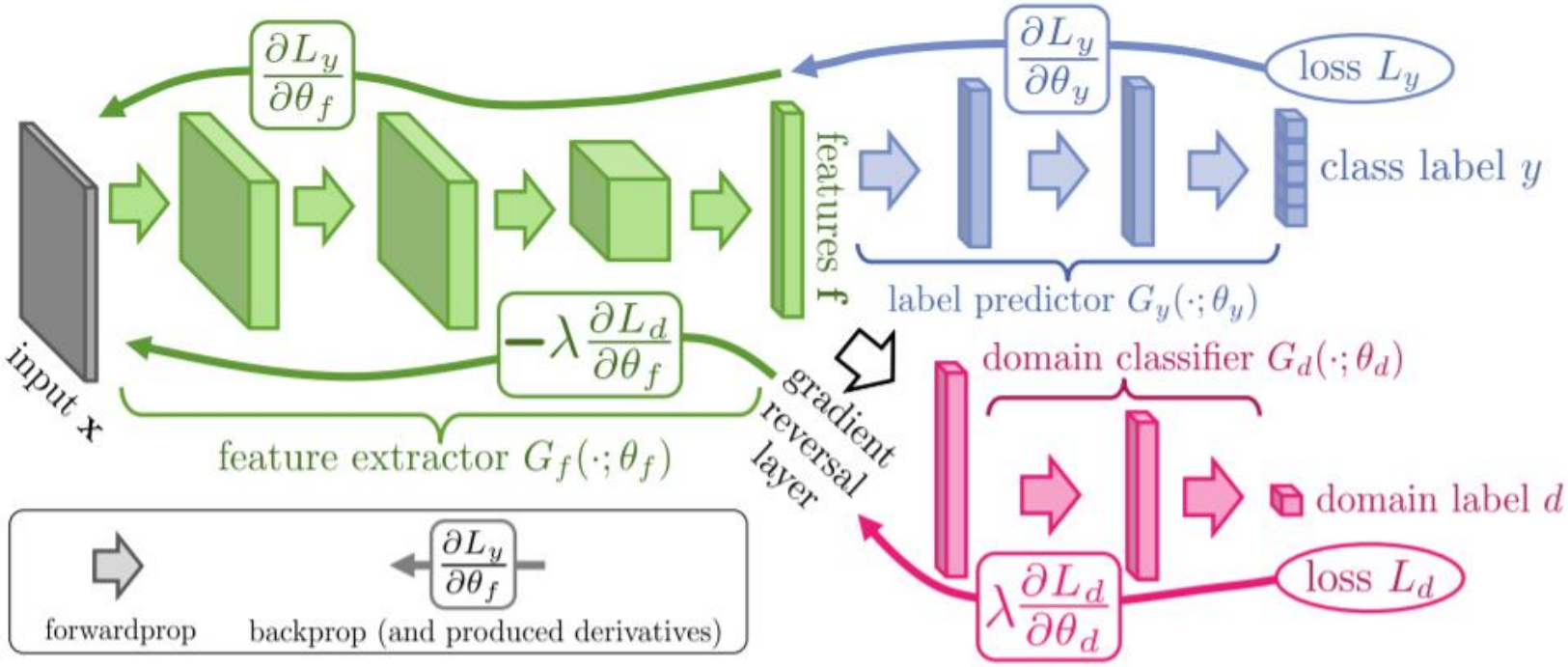
Figure 3: Entire WSIs of H&E stained tissues prepared at different facilities. Variety in staining conditions can be seen among different staining protocols.

DA training has been proposed to ignore the differences among training instances that do not contribute to the classification task.

In this paper, they consider each people to be a unique domain. Then they apply DA training approach within the MIL framework, so that the staining condition of each person's slide can be ignored.

Preliminaries: Domain-adversarial neural network

DANN Architecture:



Only 2 domains: $$\mathcal{L}_d\big(G_d(G_f(\mathbf{x}_i)), d_i\big) = d_i \log \frac{1}{G_d(G_f(\mathbf{x}_i))} + (1-d_i) \log \frac{1}{1-G_d(G_f(\mathbf{x}_i))}$$

Preliminaries: Multiscale pathology image analysis

Existing study:

(1) Use low-resolution images to find the ROI, and do high-resolution analyzation based on the ROC information.

(2) Automatically select the appropriate scale from the image itself: train another expert network to achieve adaptive scale selection.

In this study, they use multiple patches at multiple different scales simultaneously within the MIL framework, which matches the true situation: the expert pathologists conduct diagnosis by changing the magnification of a microscope repeatedly to find out various features of the tissues.

Methods

Feature extractor: $G_{\mathrm{f}} : \boldsymbol{x} \mapsto \boldsymbol{h}$

A CNN which maps a 224×224 -pixel image x into a $Q - dimensional$ feature vector h.

Bag class label predictor with attention mechanism: $G_{\mathrm{y}} : \{\boldsymbol{h}_i\}_{i \in \mathcal{I}_b} \mapsto \hat{P}(\hat{Y}_b)$

$(\boldsymbol{V}, \boldsymbol{w}) \in \theta_{\mathrm{y}}$ attention: $a_i = \frac{\exp\left(\boldsymbol{w}^{\top} \tanh(\boldsymbol{V}\boldsymbol{h}_i')\right)}{\sum_{j \in \mathcal{I}_b} \exp\left(\boldsymbol{w}^{\top} \tanh(\boldsymbol{V}\boldsymbol{h}_j')\right)}, i \in \mathcal{I}_b.$

Domain predictor: $G_{\mathrm{d}} : \boldsymbol{h} \mapsto P(\hat{d})$

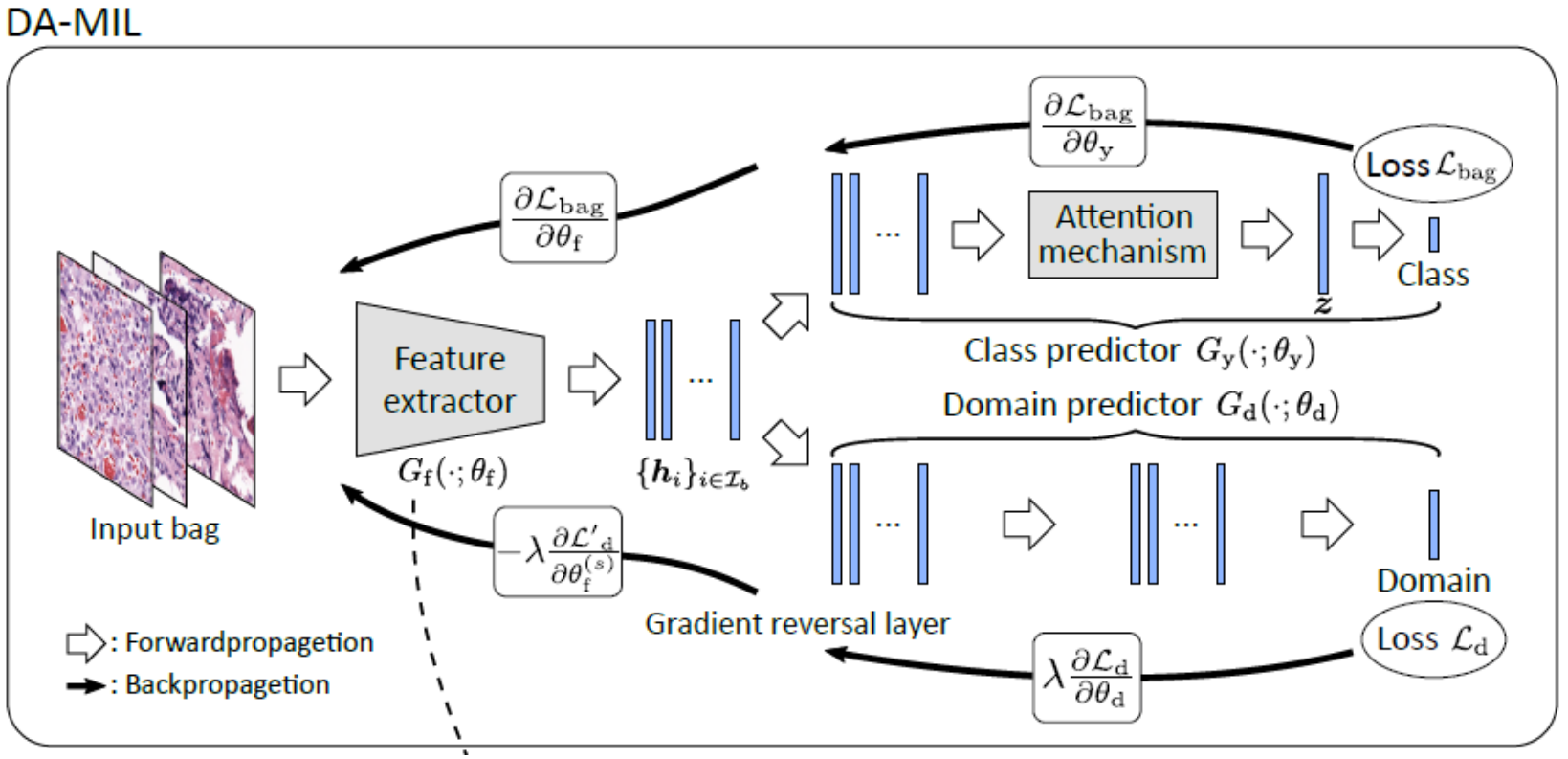
A simple NN that maps a feature vector h into domain label probabilities $P(\hat{d})$

$$p_1 = \exp\left(\frac{1}{|\mathcal{B}_n|} \sum_{b \in \mathcal{B}_n} \log P(\hat{Y}_b = 1)\right),$$

$$p_0 = \exp\left(\frac{1}{|\mathcal{B}_n|} \sum_{b \in \mathcal{B}_n} \log P(\hat{Y}_b = 0)\right).$$

$$P(\hat{\mathbb{Y}}_n = 1) = p_1/(p_1 + p_0)$$

Methods: stage 1, single scale learning



Methods: stage 1, single scale learning

$$\left(\hat{\theta}_{\mathrm{f}}^{(s)}, \hat{\theta}_{\mathrm{y}}^{(s)}, \hat{\theta}_{\mathrm{d}}^{(s)}\right) \leftarrow \arg \min_{\theta_{\mathrm{f}}^{(s)}, \theta_{\mathrm{y}}^{(s)}, \theta_{\mathrm{d}}^{(s)}} \sum_{n=1}^{N} \sum_{b \in \mathcal{B}_n} \mathcal{L}(\mathbb{Y}_n, P(\hat{Y}_b^{(s)}))$$

$$- \lambda \sum_{n=1}^{N} \sum_{b \in \mathcal{B}_n} \frac{1}{|\mathcal{I}_b^{(s)}|} \sum_{i \in \mathcal{I}_b^{(s)}} \beta_i \mathcal{L}(\mathbb{D}_n, G_{\mathrm{d}}(\boldsymbol{h}_i; \theta_{\mathrm{d}}^{(s)})), \qquad (1)$$

where

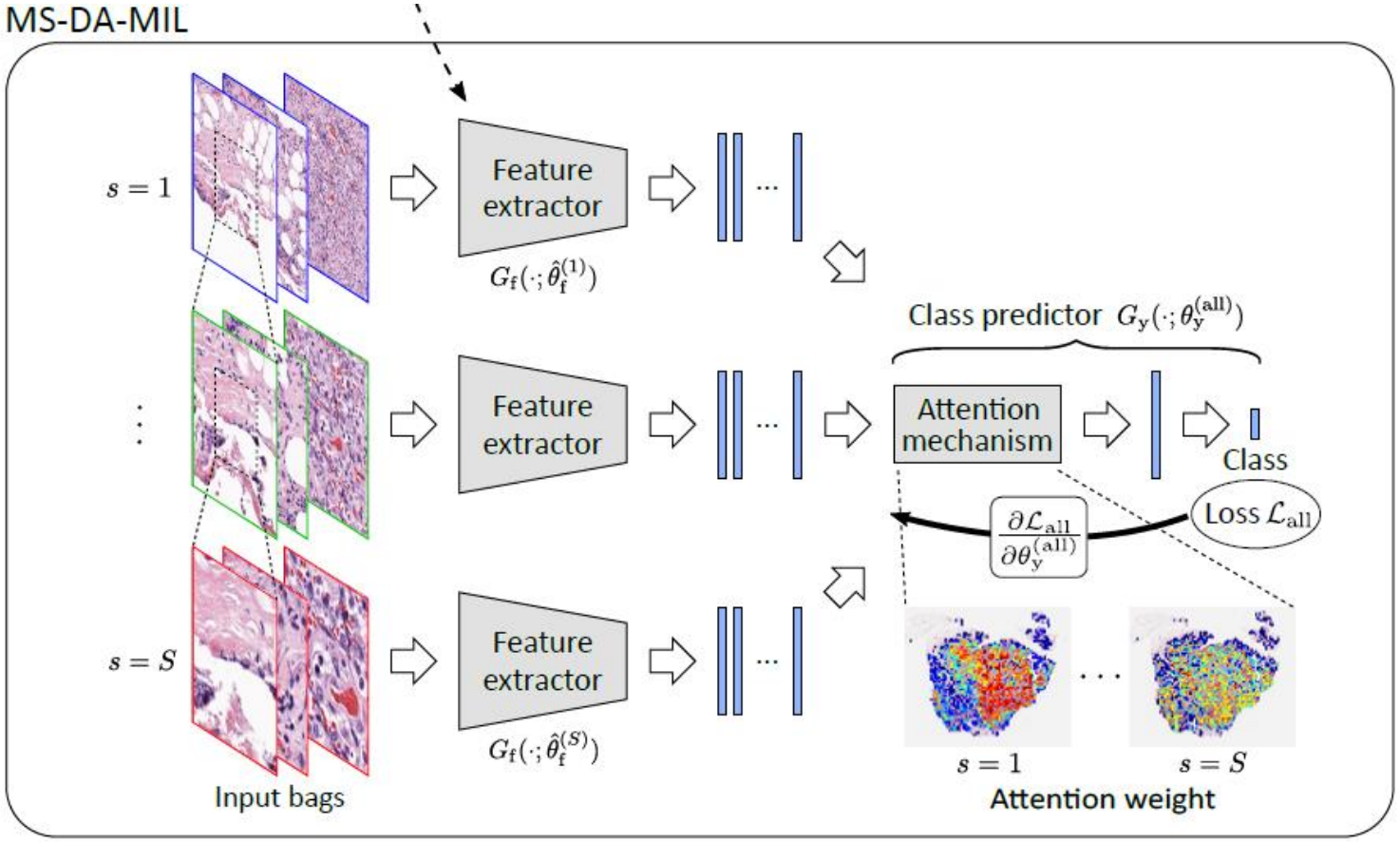
$$P(\hat{Y}_b^{(s)}) = G_{\mathrm{y}}\left(\{G_{\mathrm{f}}(\boldsymbol{x}_i; \theta_{\mathrm{f}}^{(s)})\}_{i \in \mathcal{I}_b^{(s)}}; \theta_{\mathrm{y}}^{(s)}\right),$$

$$\beta_i = \max_{a_i}\{a_j | j \in \mathcal{I}_b^{(s)}\} - a_i.$$

The first term is the loss function for bag class label prediction, while the second term is the penalty function for DA regularization, which is weighted by attentions for each instance. So the predicted bag label is more likely to be defined by the instances with larger attention.

The DA regularization term is also defined by the cross entropy between the domain label and the predicted domain label probability.

Methods: stage 2, multiple scale learning



Methods: stage 2, multiple scale learning

A multi-scale DA-MIL network is trained to predict the bag class label where each bag contains instances (patches) across different scales.

$$P(\hat{Y}_b) = G_{\mathrm{y}}\left(\{\{G_{\mathrm{f}}(\boldsymbol{x}_i;\hat{\theta}_{\mathrm{f}}^{(s)})\}_{i\in\mathcal{I}_b^{(s)}}\}_{s=1}^{S};\theta_{\mathrm{y}}^{(\mathrm{all})}\right)$$

The training of the set of parameters $\theta_y^{(all)}$ is formulated as the following minimization problem:

$$\hat{\theta}_{\mathrm{y}}^{(\mathrm{all})} \leftarrow \arg\min_{\theta_{\mathrm{y}}^{(\mathrm{all})}} \sum_{n=1}^{N} \sum_{b\in\mathcal{B}_n} \mathcal{L}(\mathbb{Y}_n, P(\hat{Y}_b))$$

Total Algorithm

Algorithm 1 Parameter update in MS-DA-MIL training.

Input: training set $\{(\mathbb{X}_n, \mathbb{Y}_n)\}_{n=1}^N$ with domain label $\{\mathbb{D}_n\}_{n=1}^N$, learning rate η, domain regularization parameter λ, train epochs M

% **stage 1**: train feature extractor $G_{\mathrm{f}}(\cdot;\theta_{\mathrm{f}}^{(s)})$, class predictor $G_{\mathrm{y}}(\cdot;\theta_{\mathrm{y}}^{(s)})$, domain predictor $G_{\mathrm{d}}(\cdot;\theta_{\mathrm{d}}^{(s)})$

for $m = 1$ to M **do**
  for $s = 1$ to S **do**
    for $n = 1$ to N **do**
      for $b = 1$ to $|\mathcal{B}_n|$ **do**

$$\{\boldsymbol{h}_i\}_{i\in\mathcal{I}_b^{(s)}} \leftarrow \{G_{\mathrm{f}}(\boldsymbol{x}_i;\theta_{\mathrm{f}}^{(s)})\}_{i\in\mathcal{I}_b^{(s)}}$$

$$\mathcal{L}_{\mathrm{bag}} \leftarrow \mathcal{L}\left(\mathbb{Y}_n, G_{\mathrm{y}}(\{\boldsymbol{h}_i\}_{i\in\mathcal{I}_b^{(s)}};\theta_{\mathrm{y}}^{(s)})\right)$$

$$\mathcal{L}_{\mathrm{d}} \leftarrow \frac{1}{|\mathcal{I}_b^{(s)}|}\sum_{i\in\mathcal{I}_b^{(s)}} \mathcal{L}\left(\mathbb{D}_n, G_{\mathrm{d}}(\boldsymbol{h}_i;\theta_{\mathrm{d}}^{(s)})\right)$$

$$\mathcal{L}'_{\mathrm{d}} \leftarrow \frac{1}{|\mathcal{I}_b^{(s)}|}\sum_{i\in\mathcal{I}_b^{(s)}} \beta_i\mathcal{L}\left(\mathbb{D}_n, G_{\mathrm{d}}(\boldsymbol{h}_i;\theta_{\mathrm{d}}^{(s)})\right)$$

$$\beta_i = \max_{a_j}\{a_j | j \in \mathcal{I}_b^{(s)}\} - a_i$$

$$\theta_{\mathrm{y}}^{(s)} \leftarrow \theta_{\mathrm{y}}^{(s)} - \eta\frac{\partial\mathcal{L}_{\mathrm{bag}}}{\partial\theta_{\mathrm{y}}^{(s)}}$$

$$\theta_{\mathrm{d}}^{(s)} \leftarrow \theta_{\mathrm{d}}^{(s)} - \eta\lambda\frac{\partial\mathcal{L}_{\mathrm{d}}}{\partial\theta_{\mathrm{d}}^{(s)}}$$

$$\theta_{\mathrm{f}}^{(s)} \leftarrow \theta_{\mathrm{f}}^{(s)} - \eta\left(\frac{\partial\mathcal{L}_{\mathrm{bag}}}{\partial\theta_{\mathrm{f}}^{(s)}} - \lambda\frac{\partial\mathcal{L}'_{\mathrm{d}}}{\partial\theta_{\mathrm{f}}^{(s)}}\right)$$

      end for
    end for
  end for
end for

% **stage 2**: train class predictor $G_{\mathrm{y}}(\cdot;\theta_{\mathrm{y}}^{(\mathrm{all})})$

for $m = 1$ to M **do**
  for $n = 1$ to N **do**
    for $b = 1$ to $|\mathcal{B}_n|$ **do**

$$\mathcal{L}_{\mathrm{all}} \leftarrow \mathcal{L}(\mathbb{Y}_n, P(\hat{Y}_b))$$

$$P(\hat{Y}_b) = G_{\mathrm{y}}(\{\{G_{\mathrm{f}}(\boldsymbol{x}_i;\theta_{\mathrm{f}}^{(s)})\}_{i\in\mathcal{I}_b^{(s)}}\}_{s=1}^S;\theta_{\mathrm{y}}^{(\mathrm{all})})$$

$$\theta_{\mathrm{y}}^{(\mathrm{all})} \leftarrow \theta_{\mathrm{y}}^{(\mathrm{all})} - \eta\frac{\partial\mathcal{L}_{\mathrm{all}}}{\partial\theta_{\mathrm{y}}^{(\mathrm{all})}}$$

    end for
  end for
end for

Output: neural network $\{\{\theta_{\mathrm{f}}^{(s)}\}_{s=1}^S, \theta_{\mathrm{y}}^{(\mathrm{all})}\}$

Experiments and results

Basic setting:196 clinical cases, which represented difficult lymphoma cases from 80 different institutions, and had been sent to an expert pathologist for diagnostic consultation.

In this experiment, we perform two-class classification, which discriminates DLBCL consisting both GCB and non-GCB types from the other three non-DLBCL classes including AITL, HLMC and HLNS.

$S = 2$: 10x (1.0um/pixel) and 20x-magnification (0.50 um/pixel), 60% training data, 20% validation data and 20% test data. In order to generate bags, 100 of 224*224-pixel image patches were randomly extracted from tissue regions in a WSI for each scale. The maximum number of bags generated from each WSI was determined as 50.

Experiments and results

Method	Magnification	Accuracy	Precision	Recall
Patch-based	10x	0.740±0.030	0.812±0.054	0.641±0.049
Patch-based	20x	0.754±0.023	0.799±0.033	0.692±0.057
Attention-based MIL	10x	0.811±0.018	0.860±0.046	0.772±0.071
Attention-based MIL	20x	0.826±0.022	0.909±0.044	0.742±0.061
DA-MIL (ours)	10x	0.836±0.012	0.927±0.037	0.743±0.046
DA-MIL (ours)	20x	0.857±0.014	0.927±0.039	0.793±0.061
MS-DA-MIL (ours)	10x, 20x	**0.871±0.028**	0.927±0.025	0.813±0.066

Experiments and results

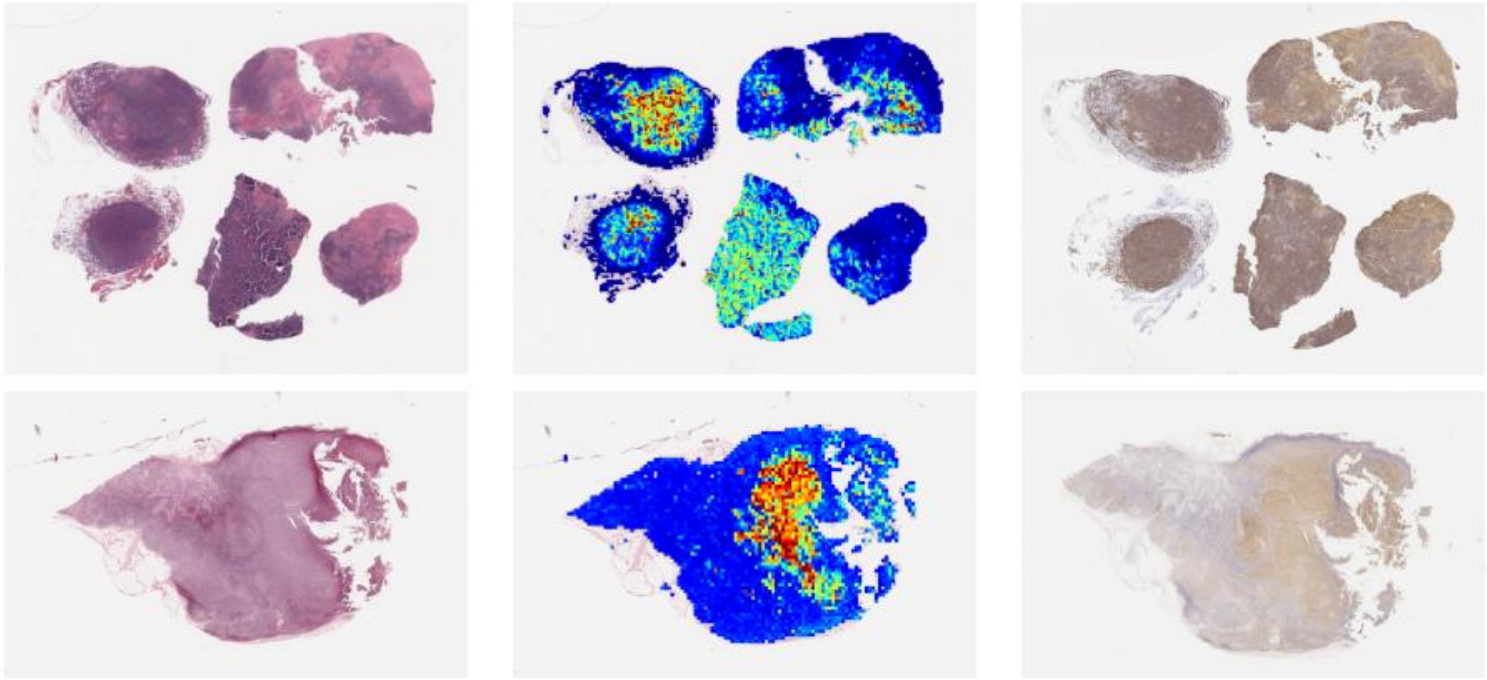


Figure 5: Visualization of attention weights in DA-MIL and corresponding IHC stained tissues: The left column is original H&E stained tissue images, the center column is visualized attention weights and the right column is CD20 stained tissue images of the same case. Attention weights in each bag are normalized between 0 to 1, and heat map from blue to red is assigned to between 0 to 1. The attention-weight map in the upper row is generated from 10x WSI, and the lower one is from 20x WSI. We can confirm that the red regions in the visualization results corresponds to the stained regions with brown in CD20 IHC stained tissue specimens.

Experiments and results

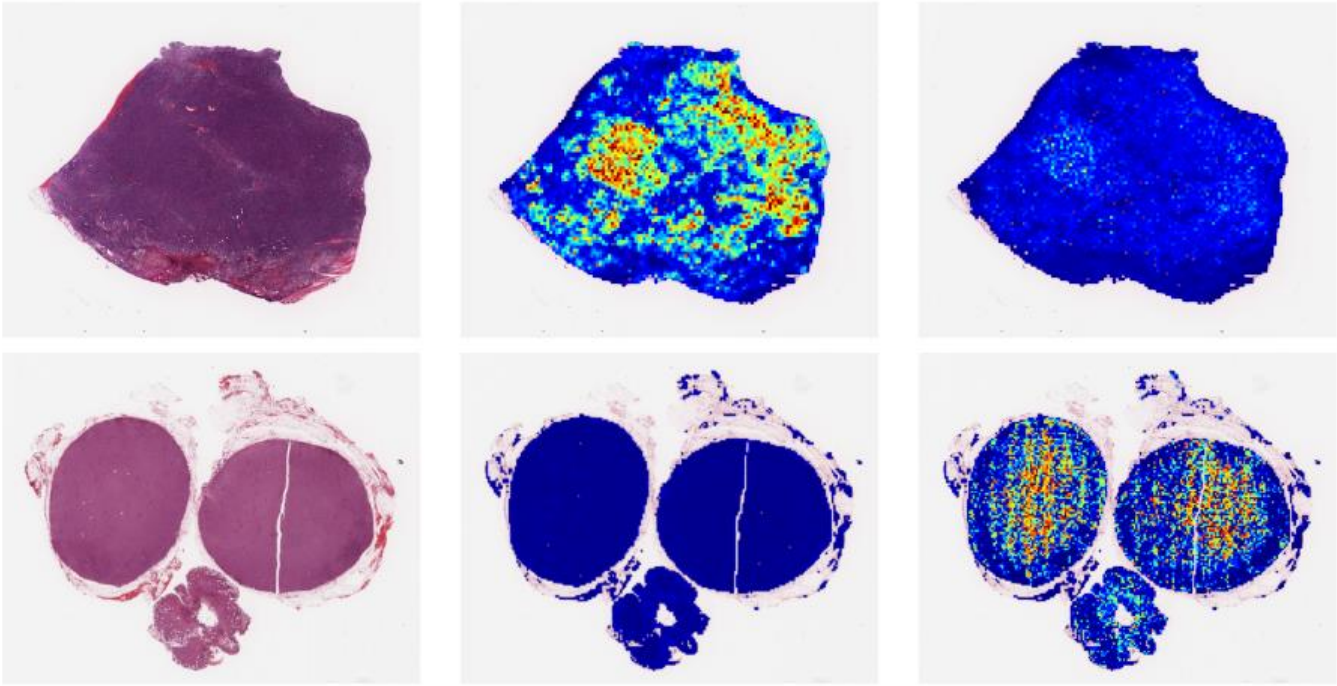


Figure 6: Visualization of attention weights in MS-DA-MIL inputs: The left column is the original H&E stained tissue images, and the center and right-hand columns are the visualized attention weights for 10x and 20x by MS-DA-MIL, respectively. We can confirm that one scale had a higher contribution for classification than the other, which means that class-specific features exist at different magnification scales depending on the individual cases.