BAYES-MIL: A NEW PROBABILISTIC PERSPECTIVE ON ATTENTION-BASED MULTIPLE INSTANCE LEARNING FOR WHOLE SLIDE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) is a popular weakly-supervised learning model on the whole slide image (WSI) for AI-assisted pathology diagnosis. The recent advance in attention-based MIL allows the model to find its region-of-interest (ROI) for interpretation by learning the attention weights for image patches of WSI slides. However, we empirically find that the interpretability of some related methods is either untrustworthy as the principle of MIL is violated or unsatisfactory as the high-attention regions are not consistent with experts' annotations. In this paper, we propose Bayes-MIL to address the problem from a probabilistic perspective. The induced patch-level uncertainty is proposed as a new measure of MIL interpretability, which outperforms previous methods in matching doctors annotations. We design a slide-dependent patch regularizer (SDPR) for the attention, imposing constraints derived from the MIL assumption, on the attention distribution. SDPR explicitly constrains the model to generate correct attention values. The spatial information is further encoded by an approximate convolutional conditional random field (CRF), for better interpretability. Experimental results show Bayes-MIL outperforms the related methods in patch-level and slide-level metrics and provides much better interpretable ROI on several large-scale WSI datasets.

1 INTRODUCTION

In real-world applications of deep learning, data like images or texts are often associated with insufficient labels, due to the expensive annotation cost. For example, the whole slide images (WSI) for medical diagnosis have about $10^5 \times 10^5$ pixels per image, but are tagged with single categorical labels (Zhang et al., 2019; Campanella et al., 2019). Weakly-supervised learning methods are designed for learning representations and making decision in these cases. Multiple instance learning (MIL) is a popular weakly-supervised learning model for the application of WSI recognition (Ilse et al., 2018; Lu et al., 2021). Concretely, a large WSI slide is sliced into a bag of image patches (instances) with a moderate size.[1] MIL builds an end-to-end parametric model that aggregates the learned features from instances and only learns from bag-level labels. The rule of aggregation is implementing *the key principle of MIL*: for binary classification, a bag is negative when all instances are negative, and a bag is positive when there is one or more positive instance (Ilse et al., 2018).

Recent advances study the *attention-based MIL* for re-weighing the instances for better performance. This attention mechanism for MIL is extensively explored and used as a measure of interpretability in various downstream tasks for medical diagnosis, like prostatic cancer (Zhang et al., 2021), breast cancer (Naik et al., 2020), etc. Specifically, the high attention weights are used to indicate that its associated instances are positive instances, e.g, the cancerous image patches. However, this rule is not formally justified and it is not clear whether the negative instances (i.e., benign) would be assigned a high attention value or the other way around. We first analyze the convergence of attention and provide validity of this rule under binary labels. Based on this rule, we conduct an empirical study on a large scale WSI dataset for how the attention mechanism in the related MIL methods performs. The study shows two clear flaws of the related methods:

[1] We define the following interchangeable terms for simplicity: "bag" and "slide"; "instance" and "patch".

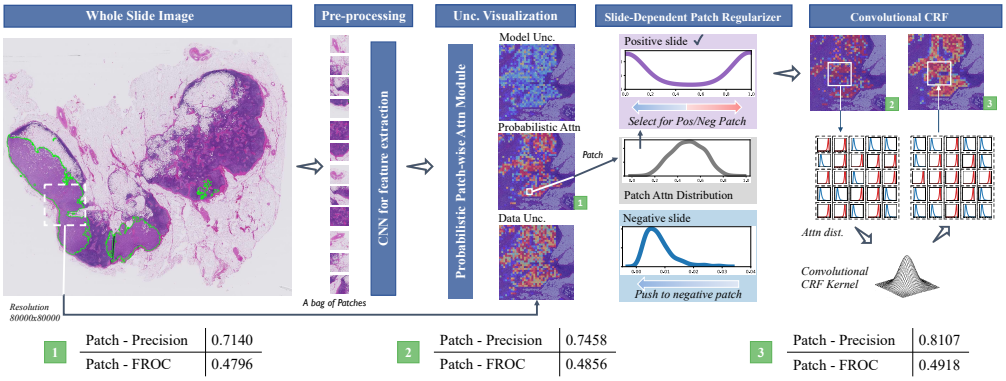


Figure 1: The overview of our Bayes-MIL framework and zoom-in views for clear visualization of interpretability. **(1)** The basic Bayes-MIL improves patch-level localization performance (Sec. 3.1). **(2)** The slide-dependent patch regularizer makes attention densely concentrated on the positive area, improving its interpretability (Sec. 3.2). **(3)** The convolutional CRF improves the localization by smoothing the uncertainty over different patches (Sec. 3.3). (bottom) The ablation results on a few metrics show the improvement of interpretability. The full ablation results are in Sec. 5.

- The interpretability for negative bags is untrustworthy because some methods violate *the key principle of MIL* by placing high attention values on negative bags, thus indicating positive instances.
- The interpretability for positive bags is unsatisfactory because the high attention values could not well match experts' annotations of positive instances.

In this paper, we address the problems from a probabilistic perspective. First, a basic framework of Bayesian MIL (Bayes-MIL) is proposed, for inducing uncertainty over the attention weights. The uncertainty is potentially an accurate measure for guessing whether the instances are positive or negative, as a replacement of attention. Second, a regularizer is designed by deduction from the MIL principle and implemented via the variational inference framework, which sets specific constraints for the attention distributions of positive and negative bags. Third, to encode the spatial information of instances for medical imaging application, we propose an approximate operation to the convolutional conditional random field, which benefits the localization of the region of interest (ROI). The final classifier is modeled in a Bayesian way, in order to provide calibrated uncertainty of the bag-level prediction. The overview of our proposed method is shown in Fig. 1. The contributions of this paper are listed as follows:

- We analyze the attention-based MIL on the interpretability-critic medical application and point out the flaws by directly using attention for interpretation.
- To address these problems, we propose the first Bayesian MIL for WSI with 3 key components: a probabilistic instance-wise attention module for uncertainty visualization, the slide-dependent patch regularizer for learning the correct attention distribution, and an approximate convolutional conditional random field for encoding spatial information. Our model provides well-calibrated uncertainties, which is crucial for safety in medical applications.
- The evaluation on large-scale MIL datasets shows Bayes-MIL outperforms the related methods in instance-level interpretation and bag-level prediction under various evaluation metrics. The visualized distribution of data uncertainty shows a strong correlation of the designed regularizer, which validates the soundness of regularizer and explains why uncertainty is useful in MIL interpretation.

2 Formulation and Analysis of Multiple Instance Learning

Multiple instance learning formulation We follow the standard formulation of Attention-based Multiple Instance Learning (MIL) (Ilse et al., 2018; Lu et al., 2021). In MIL, the input is a bag of instances, $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_K\}, \boldsymbol{x}_k \in \mathbb{R}^D$. K is the number of instances, which varies for different bags. There is a bag-level label Y. We further assume the instances also have corresponding instance-level labels $\{y_1, \dots, y_K\}$, which are *unknown* during training. There are N such bag-label pairs constituting the dataset $\mathcal{D} = \{\boldsymbol{X}_n, Y_n\}_{n=1}^N$. The objective of MIL is to learn an optimal function for predicting the bag-level label with the bag of instances as input. To this end, the MIL model should be able to aggregate the information of instances $\{\boldsymbol{x}_k\}_{k=1}^K$ to make the final decision. A well-adopted aggregation method is the embedding-based approach which maps $\boldsymbol{X}$ to a bag-level representation $\boldsymbol{z} \in \mathbb{R}^D$ and use $\boldsymbol{z}$ to predict Y. Ilse et al. (2018) extends the embedding-based

aggregation approach by leveraging the attention mechanism, namely attention-based deep MIL (ABMIL). First, a transformation $g(\cdot)$ computes a low-dimensional embedding $\boldsymbol{h}_k = g(\boldsymbol{x}_k) \in \mathbb{R}^D$ for each instance $\boldsymbol{x}_k$. The attention module aggregates the set of embeddings $\{\boldsymbol{h}_k\}_{k=1}^K$ into a bag level embedding $\boldsymbol{z}$, $\boldsymbol{z} = \sum_{k=1}^K a_k \boldsymbol{h}_k$, where the attention for the k-th instance is computed via a softmax function,

$$a_k = f_{\boldsymbol{\pi}}(\boldsymbol{H})_k = \frac{\exp\{\boldsymbol{m}^T(\tanh(\boldsymbol{V}_1^T\boldsymbol{h}_k) \odot \text{sigmoid}(\boldsymbol{V}_2^T\boldsymbol{h}_k))\}}{\sum_{j=1}^K \exp\{\boldsymbol{m}^T(\tanh(\boldsymbol{V}_1^T\boldsymbol{h}_j) \odot \text{sigmoid}(\boldsymbol{V}_2^T\boldsymbol{h}_j))\}}. \tag{1}$$

where the attention function $f_{\boldsymbol{\pi}} : \mathbb{R}^{D\times K} \to \mathbb{R}^K$ with parameters $\boldsymbol{\pi} = \{\boldsymbol{m}, \boldsymbol{V}_1, \boldsymbol{V}_2\}$, and $f_{\boldsymbol{\pi}}(\boldsymbol{H})_k$ denotes the k-th output of f. $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K] \in \mathbb{R}^{D\times K}$ is the matrix of embeddings. The bag embedding $\boldsymbol{z}$ is then mapped to the logits $\boldsymbol{u}$ with a feed forward layer with parameter $\boldsymbol{W}$ for the bag-level classification, $\boldsymbol{u} = \boldsymbol{W}^T\boldsymbol{z}$.

Multiple instance learning for medical imaging WSI is a type of high-dimensional image data format (up to $10^5 \times 10^5$ pixels per image) widely adopted in the medical area. Due to the scarcity of experts and the high annotation cost, only class labels (e.g., diagnosis results) are available for most WSIs. The high resolution of data and lack of precise annotations raise challenges for machine learning-assisted classification. To fit the data in modern computation hardware, the high-resolution image slides (*slide*) are partitioned into image patches (*patch*) before further processing. MIL fits the classification task of WSI by corresponding the slides to bags, and patches to instances. The transformation $g(\cdot)$ is the feature extractor from a pre-trained convolutional neural network. The MIL model only predicts the slide-level label, e.g., whether a slide is cancerous or not. However, the interpretation of the patches is crucial, e.g., which patches indicate the cancer, as users always check the interpretation before trusting the prediction. ABMIL uses the attention weights to tell which patches the MIL model focuses on, and several works DSMIL (Li et al., 2021), CLAM (Lu et al., 2021), TransMIL (Shao et al., 2021) study the variants of attention for better interpretability.

Ilse et al. (2018) suggested that, with binary classification label $Y \in \{0, 1\}$, the high attention weights in ABMIL could locate the positive area ($y_k = 1$) in an ideal case. The followup works from ABMIL use the high attention weights to indicate the positive patches in a black-box manner. However, there is no formal justification for this claim, and it is still unclear whether a high attention weight could also be assigned to a negative area during training. Therefore, we analyze the general attention-based MIL framework to provide this justification.

The convergence of attention Assume that the jth input patches in ith slide is $\boldsymbol{h}_{ij}$, the classifier weight is $\boldsymbol{w} \in \mathbb{R}^D$, the attention variable is $\boldsymbol{a} \in (0, 1)^{K\times 1}$. The output of the network $\hat{y}_i$ and loss function $\mathcal{L}_i$ for the i-th slide are

$$\hat{y}_i = \sigma(\boldsymbol{w}^T\boldsymbol{H}_i\boldsymbol{a} + b), \quad \mathcal{L}_i = -Y_i \log(\hat{Y}_i) - (1 - Y_i)\log(1 - \hat{Y}_i). \tag{2}$$

The $\boldsymbol{H}_i$ is all patches in ith slide, where the l_2 norm of all $h_{i,j}$ is upper bounded by 1, and $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. Assume the positive and negative patches are linear separable and there is an optimal $\boldsymbol{w}^*$, where $\|\boldsymbol{w}^*\| = 1$ and the margin is $\gamma = \min_{i,j} |\boldsymbol{w}^{*T}\boldsymbol{h}_{i,j}|$. If a slide $\boldsymbol{H}_i$ is negative, then all patches in $\boldsymbol{H}_i$ are negative, i.e., $\boldsymbol{w}^{*T}\boldsymbol{h}_{i,j} < 0, \forall j$ if $Y_i = 0$. If a slide $\boldsymbol{H}_i$ is positive, then the first K_p patches are positive and the last K_n patches are negative, i.e., $\boldsymbol{w}^{*T}[\boldsymbol{h}_{i,1}, \cdots, \boldsymbol{h}_{i,K_p}] > \boldsymbol{0}$ and $\boldsymbol{w}^{*T}[\boldsymbol{h}_{i,K_p+1}, \cdots, \boldsymbol{h}_{i,K_p+K_n}] < \boldsymbol{0}$ if $Y_i = 1$. There is an optimal $\boldsymbol{a}^*$ so that the first K_p dimension is $(1 - \epsilon)/K_p$ and the last dimension is ϵ, where ϵ is an infinitesimal, e.g., 1e-5. For the initialization of $\boldsymbol{a}$, we assume $\boldsymbol{w}^{*T}\boldsymbol{H}_i\boldsymbol{a} > \zeta > 0, \forall i$. The $\boldsymbol{a}$ is output of softmax function, $\boldsymbol{a} = s(\boldsymbol{u})$, where $a_j = \frac{\exp(u_j)}{\sum_{k=1}^K \exp(u_k)}$.

Lemma 1. *If we train the $\boldsymbol{w}$ and $\boldsymbol{a}$ as described in Appendix 1, the $\boldsymbol{w}$ converges to the optimal $\boldsymbol{w}^*$ in at most $4/max(\gamma, \zeta)^2$ steps and $\boldsymbol{a}$ converges to the desired $\boldsymbol{a}^*$, where the first K_p elements are large and the last K_n elements are small.*

Lemma 1 indicates that MIL is guaranteed to converge to the desired attention variable under the ideal condition. Note that in reality, the position of positive patches is not fixed so we use a parameterized function to make the attention variable a_j depend on the patch feature $\boldsymbol{h}_{i,j}$. Lemma 1 provides a guarantee for the validity of visualization and the design of our framework. However, in reality, the convergence of weight and attention variables depends on their initialization. Thus, existing methods may not have a good match between high-attention patches and ground-truth positive patches. The following empirical study is performed to illustrate this.

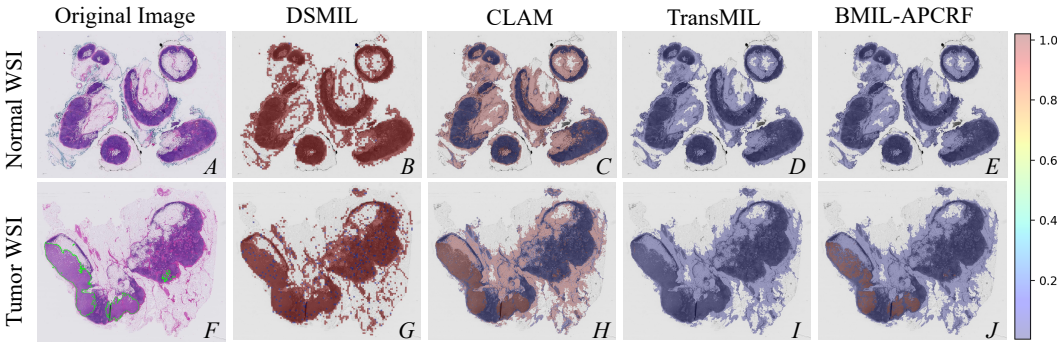


Figure 2: The visualization of normal and tumor slides and the ROIs provided by different models. The patch-level annotations for the tumor image are shown in green color in (F). The attention values $\boldsymbol{a}$ are normalized to the same range by $\frac{\boldsymbol{a}-\min_{\boldsymbol{a}}}{\max_{\boldsymbol{a}}-\min_{\boldsymbol{a}}}$. The $\min_{\boldsymbol{a}}$ and $\max_{\boldsymbol{a}}$ are the same for all methods for better visualization.

Empirical study of interpretability We use the attention $\{a_k\}_{k=1}^K$ as the patch-level prediction confidence and compare with the patch-level ground-truth $\{y_k\}_{k=1}^K$, using the related MIL approaches. One key discovery from the related methods is that *high attention values are still generated, indicating positive patches, for negative slides*, as shown in Fig. 2B 2C 2D. Another discovery is that the high attention values are not concentrated on the doctor's annotation well during inference time, as shown in Fig. 2G 2H 2I. (see numerical results in Tab. 1)

The need for Bayesian modelling of MIL The empirical study constitutes one motivation for our probabilistic approach: bring more stochasticity to the optimization process so that the convergence of weight and attention variables does not heavily depend on the initialization. Thus, we study the probabilistic counterpart of MIL, namely Bayes-MIL, which is further potentially beneficial in the following three aspects:

- *Better optimization:* Besides stochasticity in optimization, by turning $\boldsymbol{a}$ to stochastic nodes, explicit regularization could be imposed for generating correct attention (see Fig. 2E 2J).
- *New measure of interpretability:* By learning a proper posterior over the parameters $p(\boldsymbol{\pi}|\mathcal{D})$, we can induce the uncertainty over patches by $p(\boldsymbol{a}|\boldsymbol{H}^*,\mathcal{D}) = \int p(\boldsymbol{a}|\boldsymbol{\pi},\boldsymbol{H}^*)p(\boldsymbol{\pi}|\mathcal{D})d\boldsymbol{\pi}$, where $\boldsymbol{H}^*$ is the testing data. Patch-level uncertainty can potentially indicate which patches the model is uncertain about, becoming a new measure of interpretability for MIL. In other words, the patch-level uncertainty is leveraged for localizing positive areas.
- *Calibrated uncertainty:* A properly learned posterior $p(Y|\boldsymbol{H}^*,\mathcal{D})$ provides a *calibrated uncertainty* on Y, which is crucial in the application of medical imaging but ignored in existing methods.

3 Bayesian Multiple Instance Learning

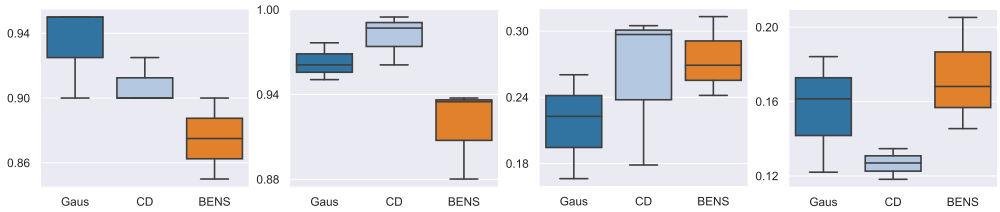
3.1 Instance-level Disentangled Uncertainty

The basic framework of Bayes-MIL is introduced in this section. To obtain uncertainty over the attention, the principled way is to assume a prior distribution $p(\boldsymbol{\pi})$ on the parameters of attention function $f_{\boldsymbol{\pi}}(\cdot)$, and let the model learn a posterior $p(\boldsymbol{\pi}|\mathcal{D})$. The other way is to directly learn an empirical posterior distribution $p(\boldsymbol{\pi}|\mathcal{D})$, e.g., ensembles. The posterior induces the uncertainty over the attention by $p(\boldsymbol{a}|\boldsymbol{H}^*,\mathcal{D}) = \int p(\boldsymbol{a}|\boldsymbol{\pi},\boldsymbol{H}^*)p(\boldsymbol{\pi}|\mathcal{D})d\boldsymbol{\pi} \approx \frac{1}{S}\sum_{s=1}^S f_{\boldsymbol{\pi}_s}(\boldsymbol{H}^*), \boldsymbol{\pi}_s \sim p(\boldsymbol{\pi}|\mathcal{D})$, where the attention function directly models the conditional distribution, i.e., $p(\boldsymbol{a}|\boldsymbol{\pi},\boldsymbol{H}) = f_{\boldsymbol{\pi}}(\boldsymbol{H}) = f(\boldsymbol{\pi},\boldsymbol{H})$.

For making $p(\boldsymbol{a}|\boldsymbol{\pi},\boldsymbol{H})$ a strict distribution, the softmax function in (1) could be leveraged for normalization, $\sum_k a_k = 1$. Then, $\boldsymbol{a}$ is a vector over the simplex, representing one categorical distribution. However, in this case, we could only calculate the uncertainty for the single categorical distribution. To extract the patch-level uncertainty, there should be one probabilistic distribution for the attention of each patch. Therefore, we need to model the distribution for patches, $p(a_k|\boldsymbol{\pi},\boldsymbol{H}) = f_{\boldsymbol{\pi}}(\boldsymbol{H})_k = f(\boldsymbol{\pi},\boldsymbol{H})_k$ and normalize for each patch. To this end, we replace the softmax function in (1) by an element-wise sigmoid function,

$$a_k = \frac{1}{1+\exp\{-\boldsymbol{m}^T(\tanh(\boldsymbol{V}_1\boldsymbol{h}_k^T)\odot\mathrm{sigmoid}(\boldsymbol{V}_2\boldsymbol{h}_k^T))\}}, \quad \boldsymbol{z} = \frac{1}{\sum_k a_k}\sum\nolimits_k a_k\boldsymbol{h}_k \tag{3}$$

where the parameters $\boldsymbol{\pi} = \{\boldsymbol{m},\boldsymbol{V}_1,\boldsymbol{V}_2\}$ are shared across all patches. The second normalization when computing $\boldsymbol{z}$ is for numerical stability.



(a) Accuracy (↑) (b) AUROC (↑) (c) NLL (↑) (d) ECE (↓)

Figure 3: The comparisons of multiplicative Gaussian, concrete dropout and BatchEnsemble for modelling probabilistic MIL weights. The slide-level results are reported. NLL is negative log-likelihood and ECE is the expected calibration error for measuring uncertainty calibration.

With this treatment, the patch-level model uncertainty and data uncertainty for each patch a_k could be extracted by (Houlsby et al., 2011; Gal et al.; 2017b):

$$\underbrace{\mathcal{I}[a_k, \boldsymbol{\pi}|\boldsymbol{H}^*, \mathcal{D}]}_{\text{Model Uncertainty}} = \underbrace{\mathcal{H}[\mathbb{E}_{p(\boldsymbol{\pi}|\mathcal{D})}[p(a_k|\boldsymbol{\pi}, \boldsymbol{H}^*)]]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{p(\boldsymbol{\pi}|\mathcal{D})}[\mathcal{H}[p(a_k|\boldsymbol{\pi}, \boldsymbol{H}^*)]]}_{\text{Data Uncertainty}} \tag{4}$$

We explore three different methods for modeling the probabilistic weights $\boldsymbol{\pi}$: Batch-Ensemble (Wen et al., 2020; Dusenberry et al., 2020) as the empirical posterior, Concrete Dropout (Gal et al., 2017a) and multiplicative Gaussian noise (Kingma et al., 2015; Molchanov et al., 2017; Cui et al., 2021) that assume prior on $\boldsymbol{\pi}$ and derive the posterior based on variational inference. As there is no direct indication of which set of posterior and prior, we should choose, we validate them empirically with the proposed framework. The experiments are conducted on CAMELYON16 dataset with 5 splits of training and validation sets. For the Batch-Ensemble, we use an ensemble of size 4. For Concrete Dropout and multiplicative Gaussian noise, we take 4 samples from the learned posterior during inference time. The multiplicative Gaussian method is selected for the probabilistic weights due to a high accuracy and a low NLL, shown in Fig. 3.

For neural networks, computing the posterior distribution using the Bayes rule requires computing intractable integrals over $\boldsymbol{\pi}$. In this paper, to be consistent with the stochastic attention function modelling, we use variational inference for approximating the posterior. Specifically, the posterior $p(\boldsymbol{\pi}|\mathcal{D})$ is approximated by $q_{\boldsymbol{\phi}}(\boldsymbol{\pi})$, by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{\pi})||p(\boldsymbol{\pi}|\mathcal{D})]$, where $\boldsymbol{\phi}$ are the variational parameters. This is equivalent to maximizing the *evidence lower bound*:

$$\max_{\phi} L_{\phi} = L_{\mathcal{D}}(\boldsymbol{\phi}) - \mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{\pi})||p(\boldsymbol{\pi})],\ L_{\mathcal{D}}(\boldsymbol{\phi}) = \sum\nolimits_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\pi})}[\log p(Y_i|\boldsymbol{H}_i, \boldsymbol{\pi})], \tag{5}$$

where $L_{\mathcal{D}}(\boldsymbol{\phi})$ is the expected data log-likelihood and $p(\boldsymbol{\pi})$ is the prior over $p(\boldsymbol{\pi})$.

3.2 Slide-Dependent Patch Regularizer

Although the proposed model can naturally visualize different types of uncertainty at the patch level and better localize the ROI, we can further *leverage the slide-level information for building a strong regularization on the ROI localization*. The aim is to explicitly encode the underlying logic of MIL into the training process, instead of letting the model explore implicit decision rules during training.

Recall Y is the slide-level label and $\{y_k\}_{k=1}^K$ are the labels for patches which are unknown during training. The intuition is based on the logic under the MIL framework that a slide is negative when all patches are negative, while a slide is positive when there are one or more positive patches. For binary label, $Y = 0$ iff $\sum_k y_k = 0$, and $Y = 1$ otherwise. The following design principle can be drawn by simple deduction: When $Y = 0$, the attention distributions $\{p(a_k|\boldsymbol{\pi}, \boldsymbol{H})\}_{k=1}^K$ must concentrate on the negative side ($a_k = 0$) to guarantee $y_k = 0$. When $Y = 1$, the attention distributions are free to select either the positive ($a_k = 1$) or negative sides. However, for precise localization, the attention distributions must concentrate on either the positive or negative side with high confidence.

This design principle is implemented by a variational inference framework by finding a regularizer on patches that is dependent on the slide label. Specifically, we choose the logit-normal distribution for a_k, due to its expressiveness over the simplex. The regularizer (a non-strict prior) is defined as

$$p(a_k|Y) = (1 - Y)\mathcal{LN}(\mu_0, \sigma_0) + Y\mathcal{LN}(\mu_1, \sigma_1), \tag{6}$$

where $\mathcal{LN}(a_k|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{a_k(1-a_k)} e^{-\frac{\mathrm{logit}(a_k)-\mu}{2\sigma^2}}$ is the logit-normal distribution. $\{\mu_0, \sigma_0\}$ and $\{\mu_1, \sigma_1\}$ are the pairs of mean and variance for the negative slides and the positive slides, respectively. As shown in Fig. 4, by setting the parameters for the regularizer, we implement the design

principle: $p(a_k|\mu_0, \sigma_0)$ concentrates on the negative side and $p(a_k|\mu_1, \sigma_1)$ allows selecting either the negative or the positive side with high confidence.

For the MIL model, we divert $\{a_k\}_{k=1}^K$ from induced distributions to stochastic nodes. The approximate posterior is defined as $q(a_k|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \mathcal{LN}(\mu_k = f_\mu(\boldsymbol{\pi}, \boldsymbol{H})_k, \sigma_k = f_\sigma(\boldsymbol{\pi}, \boldsymbol{H})_k)$ and $\sum_k \mathrm{KL}[q(a_k|\boldsymbol{\mu}, \boldsymbol{\sigma})||p(a_k|Y)]$ is used as a regularization term during training, which we denote as the slide-dependent patch regularizer (SDPR). When a negative slide is given, the mode of the attention posterior is pushed to the negative side, generating low attention values over all patches. When a positive slide is given, the posterior will be trained to select a positive mode or a negative mode, with only high concentration. This produces dense localization of the ROI. The derivation of KL divergence for the regularization is in the Appendix.

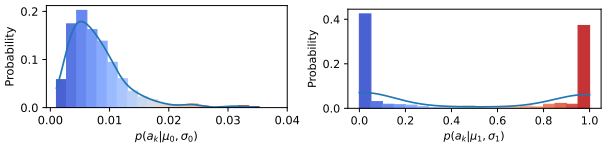


Figure 4: The visualization of density (curves) of the regularizer for the (left) negative and (right) positive slides. Samples bars are visualized with color based on the attention value.

The SDPR is also beneficial for improving the slide-level performance of MIL. The reason is it explicitly constrains the model to generate low attention values for all patches from the negative slides. Note that the major obscurity in MIL comes from the positive slides, where positive patches and negative patches coexist. There is *no obscurity* for negative slides as all patches are negative by definition. However, the previous MIL models are still free to generate high attention for patches from negative slides, which neglect the underlying logic of MIL.

3.3 Encoding Spatial Information via Approximation to Convolutional CRF

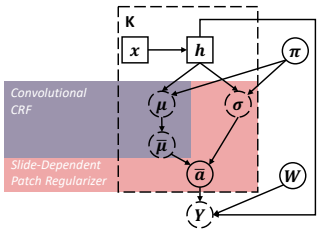


Figure 5: The graphical representation of Bayes-MIL. Boxes are deterministic nodes. Circles are stochastic nodes. Dashed circles are nodes of induced distributions.

The spatial information between patches are important for modelling of MIL for WSI recognition and localization, as there exist patches are spatially correlated (Shao et al., 2021). Here, we study how to encode the spatial information from a Bayesian perspective. The intuition is to let the neighboring attention posteriors $q(a_k)$ influence each other.

One principled method for encoding the spatial information is to use conditional random field (CRF). Assume $\bar{\boldsymbol{a}}$ is the output attention variable of CRF and $\boldsymbol{\Theta}$ is the input variable in CRF, from data. With CRF, the distribution of $p(\bar{\boldsymbol{a}}|\boldsymbol{m})$ could be directly modelled with $\boldsymbol{m}$ as the spatial features of patches (Zheng et al., 2015; Teichmann & Cipolla, 2018).

$$p(\bar{\boldsymbol{a}}|\boldsymbol{m}) = \frac{1}{Z(\boldsymbol{m})} e^{-E(\bar{\boldsymbol{a}}|\boldsymbol{m})}, \tag{7}$$
$$E(\bar{\boldsymbol{a}}|\boldsymbol{m}) = \sum\nolimits_k \psi_u(\bar{\boldsymbol{a}}_k|\boldsymbol{m}) + \sum\nolimits_{k \neq j} \psi_p(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{a}}_j|\boldsymbol{m})$$

where $\boldsymbol{m} = [\boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{a}]$ contains the coordinates of the each patch over the slide, $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_K]^T$ and $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_K]^T$. ψ_u is the unary potential that contains information only from the single patch and ψ_p is the pair-wise potential that captures the pair-wise correlation between patches. However, calculating the pair-wise correlation between patches has $\mathcal{O}(K^2)$ complexity. For efficiency, we only consider the local dependency around a_k, by setting the non-local pair-wise correlation to be 0. This is equivalent to applying a softmax normalization and a convolution operation for the provided input $\boldsymbol{a}$. We define an function for the proposed convolutional CRF, $\bar{\boldsymbol{a}} = C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{a})$. Specifically, $\tilde{\boldsymbol{a}} = \mathrm{softmax}(\boldsymbol{a})$, $\hat{\boldsymbol{a}} = \mathrm{reshape}(\boldsymbol{w}, \boldsymbol{h}, \tilde{\boldsymbol{a}})$, $\bar{\boldsymbol{a}} = \mathrm{convolution}(\hat{\boldsymbol{a}}, \mathcal{K})$, where $\mathcal{K}$ is a convolutional kernel with predefined hyperparameters. The detailed derivations and algorithms for the full convolutional CRF for BayesMIL are in the Appendix. For a precise estimation of the variable $\bar{\boldsymbol{a}}$, the following Monte-Carlo estimator is required,

$$\mathbb{E}[\bar{\boldsymbol{a}}] = \mathbb{E}_{q(\boldsymbol{a}|\boldsymbol{\mu},\boldsymbol{\sigma})}[C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{a})] \approx \frac{1}{S} \sum\nolimits_{s=1}^{S} [C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{a}_s)],\ \boldsymbol{a}_s \sim q(\boldsymbol{a}|f_\mu(\boldsymbol{\pi}, \boldsymbol{H}), f_\sigma(\boldsymbol{\pi}, \boldsymbol{H})) \tag{8}$$

The full convolutional CRF has the most promising results on real-world datasets. However, the repetitive sampling makes the training less efficient. To bypass the heavy-load of sampling, we use

a first-order Taylor expansion to approximate the expectation (see Appendix). The final format of approximation could be written as

$$q(\bar{a}_k|\bar{\boldsymbol{\mu}} = C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{\mu}), \boldsymbol{\sigma}) = \mathcal{LN}(\bar{\mu}_k = C_{\boldsymbol{w},\boldsymbol{h}}(f_\mu(\boldsymbol{\pi}, \boldsymbol{H}))_k, \sigma_k = f_\sigma(\boldsymbol{\pi}, \boldsymbol{H})_k), \tag{9}$$

for imposing slide-level patch regularizer, as well as uncertainty disentanglement and visualization.

3.4 Putting it all together

The graphical model for the Bayes-MIL is shown in Fig. 5. For the classifier with parameter $\boldsymbol{W}$, we also use a multiplicative Gaussian modelling (Kingma et al., 2015) as the posterior and posterior for calibrated uncertainty. The final objective is:

$$\max_{\phi} \quad L_{\mathcal{D}}(\boldsymbol{\phi}) - \lambda_0 R_{\boldsymbol{\pi},\boldsymbol{W}}(\phi) - \lambda_1 R_{\bar{\boldsymbol{a}}}(\phi) \tag{10}$$

$$L_{\mathcal{D}}(\boldsymbol{\phi}) = \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\pi},\boldsymbol{W})}[\log p(Y_i|\boldsymbol{H}_i, \boldsymbol{\pi}, \boldsymbol{W})], \quad R_{\boldsymbol{\pi},\boldsymbol{W}}(\phi) = \mathrm{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{\pi}, \boldsymbol{W})||p(\boldsymbol{\pi}, \boldsymbol{W})],$$

$$R_{\bar{\boldsymbol{a}}}(\phi) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{KL}[q(\bar{a}_k|C_{\boldsymbol{w},\boldsymbol{h}}(f_\mu(\boldsymbol{\pi}, \boldsymbol{H}_i)), f_\sigma(\boldsymbol{\pi}, \boldsymbol{H}_i))||p(\bar{a}_k|Y_n)]$$

where $R_{\boldsymbol{\pi},\boldsymbol{W}}(\phi)$ is KL term for the probabilistic weights and $R_{\bar{\boldsymbol{a}}}(\phi)$ is the SDPR for the attention. λ_0 and λ_1 are the trade hyperparameters for two terms respectively.

4 Related work

Attention-based multiple instance learning For improving interpretability and performance of multiple instance learning (MIL), ABMIL (Ilse et al., 2018) first introduces the attention mechanism for embedding-based MIL. DSMIL (Li et al., 2021) considers contrastive learning for feature extraction and builds the global connection between patch attentions. TransMIL (Shao et al., 2021) proposes a correlated MIL and implements it by multi-head self-attention and spatial information encoding for full global correlation. CLAM (Lu et al., 2021) extends ABMIL to the case of multiple classes and builds integrated toolbox for visualizing the uncertainty. Although the attention has been extensively adopted for the MIL intrepretability, wrong attention is still being generated for the negative bags (see Fig. 2 and Fig. 6). This is not tolerable in the application of medical imaging, where the intrepretability of model is crucial. The reason is, no methods verify the convergence of attention and impose the constraint on correcting attention from the MIL principle. In this work, these issues are carefully resolved and a new probabilistic framework is proposed.

Orthogonal works Chen et al. (2022) proposes a large-scale vision transformer solution to simultaneously learn the feature and classifier for WSI. Bayes-MIL freezes the feature extractor during training following a normal setup in MIL. Zhang et al. (2022) proposes a two-stage feature distillation MIL framework for enhancing the performance. Bayes-MIL studies the fundamental interpretability problem in the one-stage MIL framework. See Appendix for a review of uncertainty in DNNs.

5 Experiments

The proposed methods are evaluated on two standard WSI datasets: CAMELYON16 (Bejnordi et al., 2017) and CAMELYON17 (Bandi et al., 2018). CAMELYON16 contains 400 hematoxylins and eosin (H&E) stained WSI of sentinel lymph node for breast cancer, labeled as normal or tumor classes. CAMELYON17 contains 1000 WSI of the same type, labeled with normal or different stages (pN-stage) of the breast tumor. Since our paper mainly studies the interpretation of MIL, we treat these stages as one class, generating binary slide-level labels (normal and tumors). We leverage the CLAM testbed for the implementation. A ResNet-50 is used for feature extraction, consistent with previous methods. Each result is obtained with *10-fold* splits of training/validation/testing sets, which is a more thorough evaluation than previous papers. Other hyperparameters are listed in the Appendix. The codes are submitted as supplemental. In our evaluation, we consider 3 variants of BayesMIL: 1) Bayes-IL-Vis is the basic Bayesian MIL in Sec. 3.1; 2) Bayes-MIL-SDPR is the model with slide-dependent patch regularizer from Sec. 3.2; 3) Bayes-MIL-Full is the whole model, The hyperparameters are $\lambda_0 = 10^{-8}$ and $\lambda_1 = 10^{-12}$. Results on CAMELYON17 show our method is better than the CLAM baseline in Patch-localization and Slide-classification, shown in the Appendix.

Patch-level tumor region localization We first evaluate the patch-level results on tumor region localization. The tumor region localization uses the Patch-Precision, Patch-FROC and Patch-AUC

Table 1: Results on CAMELYON16: (left) Patch-level localization results using Patch-FROC (P-FROC), Patch-Precision (P-Prec.), Patch-AUROC (P-AUC); (right) Slide-level classification results using Slide-Accuracy (S-Acc.), Slide-AUROC (S-AUC) and Slide-Calibration (S-ECE).

	Patch-level			**Slide-level**		
	P-Prec. (↑)	P-FROC (↑)	P-AUC (↑)	S-Acc. (↑)	S-AUC (↑)	S-ECE (↓)
DSMIL	0.1030	0.4443	0.7719	0.8682±0.05	0.8944±0.06	0.3798±0.06
CLAM	0.6068	0.4792	0.8839	0.8650±0.06	0.9177±0.04	0.1738±0.06
CLAM-T	0.7068	0.4830	0.8884	—	—	—
TransMIL	0.1726	0.4797	0.8644	0.8837±0.02	0.9307±0.04	0.1436±0.04
Bayes-MIL-Vis	0.7140	0.4797	0.8995	0.8825±0.05	0.9164±0.06	0.1702±0.05
Bayes-MIL-SDPR	0.7458	0.4856	0.9001	0.8875±0.05	0.9432±0.05	0.1621±0.05
Bayes-MIL-APCRF	**0.8107**	**0.4919**	**0.9129**	**0.9000**±0.04	**0.9479**±0.05	**0.122**±0.04

Table 2: The patch-level localization results when using probabilistic attention values, model uncertainty and data uncertainty for our methods, evaluated on CAMELYON16.

	P-Prec. (↑)			P-FROC (↑)		
	Attn	Model Unc.	Data Unc.	Attn	Model Unc.	Data Unc.
Bayes-MIL-Vis	0.7107	0.6999	**0.7140**	0.4796	0.4788	**0.4797**
Bayes-MIL-SDPR	0.7445	0.7396	**0.7458**	0.4781	**0.4856**	0.4849
Bayes-MIL-APCRF	0.8078	0.8033	**0.8107**	0.4813	**0.4919**	0.4879

metrics. The Patch-Precision is calculated by averaging the precision of classifying the patches. Each method provides a measure (attention a_k or normalized uncertainty value $\mathcal{U}_k$) for its ROI. For calculating the precision, we compare the measure with three thresholds (0.1, 0.5, 0.9) for all methods and report the best results. Concretely, if a_k or $\mathcal{U}_k$ is greater than the threshold, the patch is predicted as positive, otherwise negative. The Patch-FROC is defined as the average sensitivity (recall) at 6 predefined false positive rates: 1/4, 1/2, 1, 2, 4 and 8 FPs per WSI (Li et al., 2021). The Patch-AUROC evaluates the averaged area under ROC over patches. The ground truth is the doctor's marking the region of the tumor.

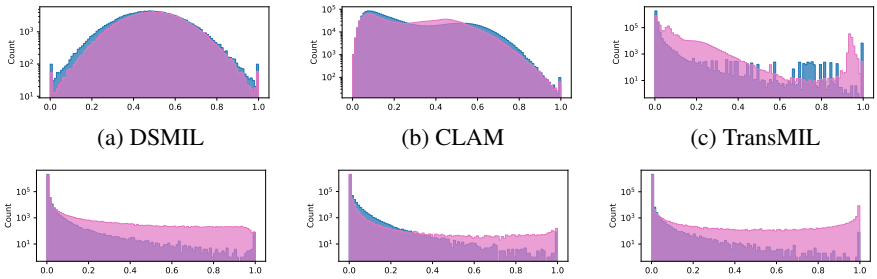
For TransMIL, we take the diagonal of attention map from the last multi-head attention module as the measure. We add another baseline, CLAM-T, on CLAM training with temperature $T = [0.2, 0.5, 2, 5]$ on softmax function, $\boldsymbol{a} = \mathrm{softmax}(\cdot, T)$, as a method for manually adjusting the density of localization, and select the best results for different metrics. For the Bayes-MIL-Vis, Bayes-MIL-SDPR and Bayes-MIL-APCRF methods, we take the best results from the normalized data uncertainty, the normalized model uncertainty and the probabilistic attention, with MC integration over 16 samples.

The results for patch-level localization of tumor regions are shown in Tab. 1 (left). The Bayesian modelling of MIL (Bayes-MIL-Vis) generally improves the patch-level visualization over other methods. Tuning the temperature for MIL attention (CLAM-T) could marginally improve localization results, however, it requires exhaustively tuning the temperature. For Bayes-MIL, the precision is improved by a large margin (by 0.1072-0.611), showing the advantage of using uncertainty for localizing positive area. Including SDPR and the approximate CRF improve the patch-level results consistently, and the full model Bayes-MIL-APCRF achieves the best performance on the 3 metrics.

Tab. 2 shows the detailed results with probabilistic attention, model uncertainty and data uncertainty for the three proposed methods. The probabilistic attention benefits from the Bayes-MIL model, making it competitive at tumor region localization. The best results for precision are from the disentangled data uncertainty, indicating the data uncertainty is able to precisely localize the tumor region. The reason is that data uncertainty captures the rareness of features of positive patches indicating anomalies, such as cancerous cells, in the whole dataset. The model uncertainty is better on the P-FROC metric, which evaluates the recall at different false positive rates. This shows that the uncertainty induced from parameter distributions in $\boldsymbol{\pi}$ tends to cover the positive patches better.

Slide-level evaluation We next evaluate the slide-level classification performance. Tab. 1 (right) shows the slide-level performance on CAMELYON16, measured by Accuracy, AUC, and ECE. The results of DSMIL and TransMIL are from their papers, as their reported performance is better than our reimplementation. Compared with DSMIL and CLAM, Bayes-MIL-Vis has a higher accuracy, a lower calibration error and a similar AUC. When including our SDPR, Bayes-MIL-SDPR has better accuracy and AUC than TransMIL.

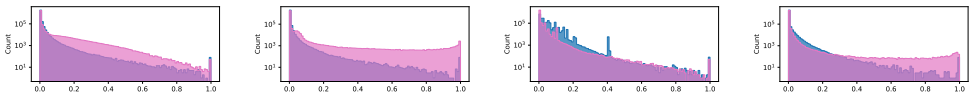
Why is Bayes-MIL good at slide classification and ROI localization? To understand the reason behind the good performance of the proposed framework, we visualize attention distribution in Fig. 6. As shown in Fig. 6a and Fig. 6b, DSMIL and CLAM generate similar attention distributions for positive and negative slides. For the negative slides, DSMIL still generates a nearly normal



(a) DSMIL (b) CLAM (c) TransMIL

(d) Bayes-MIL-Vis Attn (e) Bayes-MIL-SDPR Attn. (f) Bayes-MIL-APCRF Attn.

Figure 6: The log-scale histogram of attention for different methods. The pink and blue curves are for positive and negative slides, respectively. All testing images in CAMELYON16 are used, which contain 2M positive and 2M negative patches.



(a) BMIL-Vis M-Unc. (b) BMIL-Vis D-Unc. (c) BMIL-SDPR M-Unc. (d) BMIL-SDPR D-Unc.

Figure 7: The log-scale histogram of normalized uncertainty for Bayes-MIL. M and D stand for model and data. APCRF has similar histograms as SDPR, thus not shown for saving space.

distribution of attention, which deviates from the MIL principle. For positive slides, the densities on the positive side ($a \approx 1$) of DSMIL and CLAM are low. Note that the vanilla version BMIL-Vis has different attention distributions for positive and negative slides, but does not generate high density on the positive side for positive slides. Bayes-MIL with SDPR is able to push the attention of negative slides to the negative side ($a \approx 0$, the blue curve in Fig. 6e), while generating a U-shape distribution for the positive slides (red curve in Fig. 6e), which corresponds to our design of SDPR with the ideal case shown in Fig. 4. This design benefits the following aspects:

- For visualization, Bayes-MIL generates concentrated high attention values for positive slides, while only generates correct values ($a \approx 0$) for negative slides.
- For classification, learning correct attention guided by SDPR for the negative slides will benefit the classification of patches in the positive slides. TransMIL captures this attention distribution to some degree, so it performs well in slide-level classification.

Bayes-APCRF performs a smoothing operation on the mean of the stochastic attention nodes. This pushes the histogram to have a more distinct U-shape (the red U-shape curve in Fig. 6f) by aggregating over neighboring positive patches, which benefits the localization and obtains the best visualization scores in Tab. 1.

Correctly Visualized ROI Fig. 2 shows the visualization results on the negative (normal) and positive (tumor) WSI. The related methods all generate high attention for the negative slides, which is against of MIL principle. TransMIL only generates a small area of high attention for negative slides, however the performance on positive slides is poor. *Bayes-MIL is the only method that generate correct values ($a \approx 0$) for the negative slides, while performing the best in localization on positive slides.*

Measure of uncertainty The measure of data uncertainty (Fig. 8b and Fig. 7d) naturally captures the information from the patch distribution. It shows a similar distribution with the attention of Bayes-MIL-SDPR and the ideal case of SDPR in Fig. 4. This might be the reason why data uncertainty has the best performance in localizing the ROI. Furthermore, based on this observation on data uncertainty, the soundness of SDPR is empirically validated.

6 Conclusion

This paper analyzes the interpretability problem in existing attention-based multiple instance learning (MIL) models. Directly taking attention as a measure of the important instances empirically violates the MIL principle. To address this problem, we propose a probabilistic solution Bayes-MIL that provides new measures for interpretability. To ensure the validity of interpretation in MIL, a regularizer on attention is required. A slide-dependent patch regularizer is proposed for imposing explicit constraints to let the model learn under the MIL principle, which also improves the slide-level performance. The spatial information is further encoded, which improves both slide-level and patch-level performance. The analysis and visualization of data uncertainty distribution further validate the main idea and soundness of SDPR.

REFERENCES

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

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*, 318(22):2199–2210, 2017.

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.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022.

Yufei Cui, Wuguannan Yao, Qiao Li, Antoni B Chan, and Chun Jason Xue. Accelerating monte carlo bayesian inference via approximating predictive uncertainty over simplex. *arXiv preprint arXiv:1905.12194*, 2019.

Yufei Cui, Ziquan Liu, Qiao Li, Antoni B Chan, and Chun Jason Xue. Bayesian nested neural networks for uncertainty calibration and adaptive compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2392–2401, 2021.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pp. 2782–2792. PMLR, 2020.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017a.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017b.

Yarin Gal et al. Uncertainty in deep learning.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

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

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*, 2019.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, pp. 2498–2507. PMLR, 2017.

Nikhil Naik, Ali Madani, Andre Esteva, Nitish Shirish Keskar, Michael F Press, Daniel Ruderman, David B Agus, and Richard Socher. Deep learning-enabled breast cancer hormonal receptor status determination from base-level h&e stains. *Nature communications*, 11(1):1–8, 2020.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Marvin TT Teichmann and Roberto Cipolla. Convolutional crfs for semantic segmentation. *arXiv preprint arXiv:1805.04777*, 2018.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

Jingwei Zhang, Ke Ma, John Van Arnam, Rajarsi Gupta, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. A joint spatial and magnification based attention framework for large scale histopathology classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3776–3784, 2021.

Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1529–1537, 2015.

A PROOF OF LEMMA 1

We initialize the $\boldsymbol{w}$ as $\mathbf{0}$ and first update $\boldsymbol{w}$ for T steps, where each step uses one slide to compute the loss and gradient. In the optimization for $\boldsymbol{w}$, we use a learning rate of η and if the prediction for a slide is right, i.e., $\hat{Y}_i > 0.5$ for a positive patch or $\hat{Y}_i < 0.5$ for a negative patch, the η is set as 0. After we update the $\boldsymbol{w}$, the attention variables are optimized.

proof. The gradient for $\boldsymbol{w}$ and $\boldsymbol{a}$

$$\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}} = \boldsymbol{H}_i \boldsymbol{a}(\hat{Y}_i - Y_i), \tag{11}$$

$$\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{a}} = \boldsymbol{H}_i^T \boldsymbol{w}(\hat{Y}_i - Y_i), \tag{12}$$

$$\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{u}} = (\boldsymbol{I}_{\boldsymbol{a}} - \boldsymbol{A}^2)\boldsymbol{H}_i^T \boldsymbol{w}(\hat{Y}_i - Y_i), \tag{13}$$

where $\boldsymbol{I}_{\boldsymbol{a}} = \boldsymbol{I}\boldsymbol{a}$ and $\boldsymbol{A}^2 = [\boldsymbol{a} \odot \boldsymbol{a}, \cdots, \boldsymbol{a} \odot \boldsymbol{a}] \in \mathbb{R}^{K \times K}$.

Consider the inner product of $\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}}$ and the optimal $\boldsymbol{w}^*$,

$$(\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}})^T \boldsymbol{w}^* = \boldsymbol{w}^T \boldsymbol{w}^* - (\hat{Y}_i - Y_i)\boldsymbol{a}^T \boldsymbol{H}_i^T \boldsymbol{w}^*. \tag{14}$$

If $Y_i = 0$, $\boldsymbol{H}_i^T \boldsymbol{w}^* < \mathbf{0}$, we have

$$(\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}})^T \boldsymbol{w}^* \geq \boldsymbol{w}^T \boldsymbol{w}^* + \eta\gamma\hat{Y}_i \geq \boldsymbol{w}^T \boldsymbol{w}^* + \frac{1}{2}\eta\gamma \tag{15}$$

If $Y_i = 1$, we have,

$$(\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}})^T \boldsymbol{w}^* = \boldsymbol{w}^T \boldsymbol{w}^* + \eta(1 - \hat{Y}_i)\boldsymbol{a}^T \boldsymbol{H}_i \boldsymbol{w}^* \geq \boldsymbol{w}^T \boldsymbol{w}^* + \eta\zeta(1 - \hat{Y}_i) \geq \boldsymbol{w}^T \boldsymbol{w}^* + \frac{1}{2}\eta\zeta. \tag{16}$$

Next consider the scale of $\boldsymbol{w}$,

$$(\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}})^T (\boldsymbol{w} - \eta \frac{\partial \mathcal{L}_i}{\partial \boldsymbol{w}}) = \boldsymbol{w}^T \boldsymbol{w} - 2\eta(\hat{Y}_i - Y_i)\boldsymbol{a}^T \boldsymbol{H}_i^T \boldsymbol{w} + \eta^2(\hat{Y}_i - Y_i)^2 \boldsymbol{a}^T \boldsymbol{H}_i^T \boldsymbol{H}_i \boldsymbol{a} \tag{17}$$

$$\leq \boldsymbol{w}^T \boldsymbol{w} + \eta^2 \tag{18}$$

After T steps of update, since the $\boldsymbol{w}$ is initialized as $\mathbf{0}$, we have

$$\boldsymbol{w}^{*T} \boldsymbol{w} \geq \frac{1}{2}\eta T \max(\gamma, \zeta), \|\boldsymbol{w}\| \leq \eta\sqrt{T} \tag{19}$$

$$\frac{\boldsymbol{w}^{*T} \boldsymbol{w}}{\|\boldsymbol{w}\|\|\boldsymbol{w}^*\|} \geq \frac{\frac{1}{2}\eta T \max(\gamma, \zeta)}{\eta\sqrt{T}} = \frac{1}{2}\sqrt{T}\max(\gamma, \zeta) \tag{20}$$

As a result of the Cauchy-Schwartz inequality,

$$\frac{\boldsymbol{w}^{*T} \boldsymbol{w}}{\|\boldsymbol{w}\|\|\boldsymbol{w}^*\|} \leq 1, \tag{21}$$

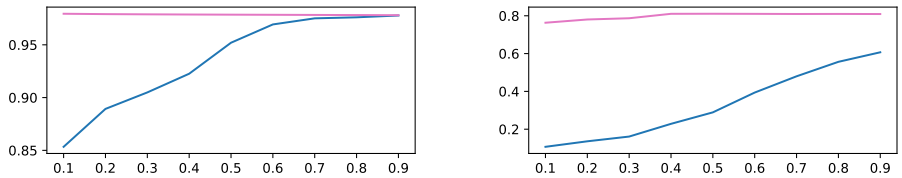
$$T \leq \frac{4}{\max(\gamma, \zeta)^2}. \tag{22}$$

Next look at the update for $\boldsymbol{a}$,

$$\frac{\partial \mathcal{L}_i}{\partial \boldsymbol{u}} = (\boldsymbol{I}_{\boldsymbol{a}} - \boldsymbol{A}^2)\boldsymbol{H}_i^T \boldsymbol{w}(\hat{Y}_i - Y_i) \approx \boldsymbol{I}_{\boldsymbol{a}} \boldsymbol{H}_i^T \boldsymbol{w}_T(\hat{Y}_i - Y_i) \tag{23}$$

$$= [a_1 \boldsymbol{h}_{i,1}^T \boldsymbol{w}_T, \cdots, a_K \boldsymbol{h}_{i,K}^T \boldsymbol{w}_T](\hat{Y}_i - Y_i). \tag{24}$$

We know that $\boldsymbol{w}_T$ converges to the optimal $\boldsymbol{w}^*$ so the $\boldsymbol{h}_{i,j}^T \boldsymbol{w}_T > 0$ if the jth patch is positive and $\boldsymbol{h}_{i,j}^T \boldsymbol{w}_T < 0$ if jth patch is negative. so this update means that the first K_p values of $\boldsymbol{u}$ will increase for positive slides while the last K_n values will decrease for any slides. This iteration leads to the desired optimal $\boldsymbol{a}^*$.



(a) Patch-level accuracy. (b) Patch-level precision.

Figure 8: The patch-level performance of positive slides of CAMELYON16 under different decision thresholds $\{0.1, 0.2, \ldots, 0.9\}$. The blue line is CLAM and the red line is Bayes-MIL.

B EXPERIMENTS

Hyper-parameters For the backbone MIL model, we use the the same parameter setup as CLAM. We use Adam with learning rate 10^{-4} and weight decay 10^{-5} for optimization. We use early stop with patient step 20 and maximum epoch 200.

Efficiency The running time information of different methods, tested on Camelyon16, are shown in Tab. 3. By using an efficient variational inference design and implementation, we keep a decent running time of Bayes-MIL between CLAM and TransMIL.

Table 3: Running time of the MIL methods.

	CLAM	TransMIL	DSMIL	Bayes-MIL
training time (minute per epoch)	0.606	1.8	0.824	1.217
testing time (second per slide)	0.062	5.857	0.045	0.185

Supplemental results on CAMELYON16 The full CRF (8) generate the best slide-level results on CAMELYON16. The incomplete results with 3-fold splits of the dataset show the testing accuracy is 0.9015, the testing AUROC is 0.9546. As the computation efficiency is low, we stopped running the experiments.

Results on normal slides of CAMELYON16 The numerical results (accuracy) on Camelyon16 are shown in Tab. 4. Bayes-MIL consistently presents high accuracy while other methods (e.g., CLAM) do not provides satisfactory results.

Table 4: Patch-level results on normal slides of CAMELYON16

Methods	Acc (threshold=0.1)	Acc (threshold=0.3)	Acc (threshold=0.5)
DSMIL	0.0058	0.1188	0.5547
CLAM	0.2758	0.6650	0.8395
TransMIL	0.9846	0.9898	0.9920
Bayes-MIL-APCRF	**0.9986**	**0.9996**	**0.9998**

Results on tumor slides of CAMELYON16 The comparisons of patch-level performance between CLAM and the data uncertainty of Bayes-MIL-APCRF are shown in Fig. 8. It could be concluded that Bayes-MIL presents consistently decent performance under different threshold while CLAM should explore its optimal threshold.

More results on CAMELYON17 The patch-level and slide-level results on CAMELYON17 are shown in Tab. 5. The ablation study shows the effectiveness of the three step model design. Bayes-MIL has better performance than the CLAM baseline.

Note that BMIL-APCRF has a slightly lower slide-level performance in terms of accuracy and AUC, but still competitive compared with previous works. The reason could be that in order to utilize the efficient standard computation library, $\tilde{\boldsymbol{a}}$ is reshaped to a $W \times F$ rectangle. As a result, some areas are 0 in the rectangle, which might introduce ambiguity for classification. In future work, we will implement an efficient convolution for the irregular feature map to improve the performance, which is orthogonal to this work.

Table 5: Patch-level and Slide-level results on CAMELYON17

	Patch-level			**Slide-level**		
	P-Prec. (↑)	P-FROC (↑)	P-AUC (↑)	S-Acc. (↑)	S-AUC (↑)	S-ECE (↓)
CLAM	0.3833	0.4454	0.8121	0.8055±0.03	0.8222±0.03	0.2589±0.01
Bayes-MIL-Vis	0.3875	0.4545	0.8222	0.8088±0.02	0.8276±0.05	0.2547±0.02
Bayes-MIL-SDPR	0.4001	0.4591	0.8265	**0.8185**±0.03	**0.8612**±0.01	0.2455±0.01
Bayes-MIL-APCRF	**0.4131**	**0.4600**	**0.8312**	0.8180±0.03	0.8597±0.02	**0.2453**±0.02

More results on TCGA-NSCLC Bayes-MIL uses $p(a_k|Y=1) = \mathcal{LN}(\mu_1, \sigma_1)$ as the prior for the tumor subtype classification, as no negative slides are provided. Results in Tab. 6 shows Bayes-MIL outperforms existing methods in accuracy and calibration, while providing AUC higher than CLAM baseline. Note this result is preliminary without search for optimal setup. Future work will explore more detailed setups on the tumor subtype classification dataset.

Table 6: Tumor subtype classification results on TCGA-NSCLC.

Methods	AUC	Accuracy	ECE
CLAM	0.9420 ±0.03	0.8640 ±0.04	0.1697±0.02
Scaling ViT	0.9516	0.8821	N/A
TransMIL	**0.9603**	0.8835	N/A
Bayes-MIL-Enc	0.9440±0.01	0.8893±0.04	0.1647±0.01
Bayes-MIL-APCRF	0.9451±0.01	**0.8966**±0.01	**0.1575**±0.02

Tumor stage classification results on CAMELYON17. Bayes-MIL uses $p(a_k|Y=1) = \mathcal{LN}(\mu_1, \sigma_1)$ as the prior for different stages of the tumors, while uses $p(a_k|Y=0) = \mathcal{LN}(\mu_0, \sigma_0)$. This is consistent with the original design for the SDPR. A softmax function is adopted for the slide level classification.

Table 7: Tumor stage classification results on CAMELYON17.

Methods	AUC	Accuracy	ECE
CLAM	0.7803	0.6	0.4138
Bayes-MIL-APCRF	**0.8070**	**0.64**	**0.4017**

C THE PSEUDO CODES FOR TRAINING AND INFERENCE

The training algorithm is as follows:

$\mathbf{H} \longleftarrow \text{feature_extractor}(\mathbf{X})$
while not converged **do**
  $\boldsymbol{\pi}, \mathbf{W} \sim q_\phi(\boldsymbol{\pi}, \mathbf{W})$.
  Calculate a_k with (3).
  $\bar{a}_k = \text{APCRF}(a_k)$.
  Calculate $L_{\mathcal{D}}(\boldsymbol{\phi})$, $R_{\boldsymbol{\pi},\boldsymbol{W}}(\phi)$, $R_{\bar{\boldsymbol{a}}}(\phi)$ and optimize with the objective (10).
end while

The inference algorithm includes an Monte-Carlo integration which takes S Monte-Carlo samples:

$\mathbf{H} \longleftarrow$ feature_extractor$(\mathbf{X})$
for s=1:S **do**
  $\boldsymbol{\pi}_s \sim q_\phi(\boldsymbol{\pi})$.
  Calculate a_{ks} with (3).
end for
$a_k = \frac{1}{S}\sum_s a_{ks}$
$\bar{a}_k = \text{APCRF}(a_k)$.
$\hat{y}_i = \sigma(\boldsymbol{w}^T \boldsymbol{H}_i \bar{\boldsymbol{a}} + b)$.

D Details on variational inference

We use the multiplicative Gaussian with log-uniform prior proposed by Kingma et al. (2015) for the MIL parameters, as the variational parameters of posterior mean could be canceled in KL term for easier optimization. Molchanov et al. (2017) revised the formulation of KL term by an approximation for sparsity of neural network. Cui et al. (2021) further extends the approximation for KL to have a flexible range. Specifically, the prior $p(\log|\pi|) = \text{const}$. The posterior is

$$\pi = \theta\eta = \theta(1+\sqrt{\phi}\epsilon_\pi),\ \epsilon_\pi \sim \mathcal{N}(0,1), \quad \pi \sim \mathcal{N}(\pi|\theta, \phi\theta^2). \tag{25}$$

By taking a mean-field approximation, the KL term is

$$-D_{\text{KL}}[q(\pi|\theta,\phi)||p(|\pi|)] = \frac{1}{2}\log\alpha - \mathbb{E}_{\epsilon_\pi\sim\mathcal{N}(1,\phi)}\log|\epsilon_\pi| + C, \tag{26}$$

For a flexible range of α, the KL term is approximated by

$$R_{\boldsymbol{\pi},\boldsymbol{W}}(\phi) = \text{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{\pi},\boldsymbol{W})||p(\boldsymbol{\pi},\boldsymbol{W})] = \sum_{\phi\in\Phi} b_1 e^{-e^{b_4\cdot(b_2+b_3*\log\phi)^2}} - 0.5\log(1+\phi^{-1}) + C, \tag{27}$$

where $b_1 = 0.7294$, $b_2 = -0.2041$, $b_3 = 0.3492$ and $b_4 = 0.5387$. Φ is the whole set of variational parameters corresponding to $\{\boldsymbol{\pi}, \mathbf{W}\}$. Refer to Molchanov et al. (2017) and Cui et al. (2021) for more details.

As Y is a hard variable of 0 or 1 during training process, (6) only chooses one component for each input slide. The KL term of SDPR is analytically written as

$$R_{\bar{\boldsymbol{a}}}(\phi) = \sum_k \log\frac{\sigma'}{\bar{\sigma}_k} + \frac{\bar{\sigma}_k^2 + (\bar{\mu}_k - \mu')^2}{2\bar{\sigma}_k^2} - 0.5 \tag{28}$$

where $\bar{\mu}_k = C_{\boldsymbol{w},\boldsymbol{h}}(f_\mu(\boldsymbol{\pi},\boldsymbol{H}_i))_k, \bar{\sigma}_k = f_\sigma(\boldsymbol{\pi},\boldsymbol{H}_i))_k$ are conditioned on the variational parameters. μ' and σ' are based on the slide label, for selecting between $\{\mu_0,\sigma_0\}$ and $\{\mu_1,\sigma_1\}$.

E Convolutional CRF

Following the notation in Sec. 3.3, the pair-wise potential function is defined as

$$\psi_p(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{a}}_j|\boldsymbol{m}) = c(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{a}}_j)\sum_{l=1}^{L}\gamma^{(l)}\mathcal{K}^{(l)}(\boldsymbol{m}_i, \boldsymbol{m}_j) \tag{29}$$

the mean-field approximation Q to conditional random field (Zheng et al., 2015) can be written as

$Q_k \leftarrow \frac{1}{Z}e^{\psi_u(\bar{\boldsymbol{a}}_k|\boldsymbol{m})}$
while not converged **do**
  1. $Q_k \leftarrow \sum_{k\neq j}\psi_p(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{a}}_j|\boldsymbol{m})$
  2. $Q_k \leftarrow \psi_u(\boldsymbol{a}_k|\boldsymbol{m}) + Q_k$
  3. $Q_k \leftarrow \text{normalize}(Q_k)$
end while

We omit the label compatibility step as $a_k \in R$ is a continuous value instead of categorical variable. The last normalization step could also be omitted as we perform normalization on the output attention in (3). Thus, the cross-patch label compatibility $c(\bar{\boldsymbol{a}}_k, \bar{\boldsymbol{a}}_j)$ can also be a constant, which we set to be 1. $w^{(l)}$ is a trainable parameter and $\mathcal{K}^{(l)}$ is a Gaussian kernel, both indexed by l. L is the total

number of kernels. We perform this algorithm in a online fashion thus the loop is only executed once in every training iteration.

As shown in Teichmann & Cipolla (2018), if we only consider local dependency, the kernel matrix could be defined as

$$\mathcal{K}(d\omega, d\eta, \omega, \eta) = exp(-\frac{|\vartheta(\omega,\eta) - \vartheta(\omega - d\omega, \eta - d\eta)|^2}{2\beta_k^2}), \tag{30}$$

where β_i is a learnable parameter. ϑ is the feature used for the kernel. A merged kernel matrix could be defined as $\kappa = \sum_l \gamma_l \mathcal{K}_k$. The message passing step (1) could be written as $Q(\omega, \eta) = \sum_{d\omega, d\eta} \mathcal{K}(d\omega, d\eta, \omega, \eta) \cdot Q(\omega + d\omega, \eta + d\eta)$. This is equivalent to a convolution operation (Teichmann & Cipolla, 2018). Thus, the step is simplified into the following function $\bar{\boldsymbol{a}} = C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{a})$. Specifically, $\tilde{\boldsymbol{a}} = \mathrm{softmax}(\boldsymbol{a})$, $\hat{\boldsymbol{a}} = \mathrm{reshape}(\boldsymbol{w}, \boldsymbol{h}, \tilde{\boldsymbol{a}})$, $\bar{\boldsymbol{a}} = \mathrm{convolution}(\hat{\boldsymbol{a}}, \mathcal{K})$. In this paper, we further take a non-parametric 3×3 convolutional kernel over the input, which is a Gaussian smoothing operation over the attention.

Next we use first-order Taylor expansion to further simplify the expectation in (8). If we perform the Taylor-expansion at the expectation of $\mathbf{a}$:

$$\mathbb{E}[\bar{\boldsymbol{a}}] = \mathbb{E}_{q(\boldsymbol{a}|\boldsymbol{\mu},\boldsymbol{\sigma})}[C_{\boldsymbol{w},\boldsymbol{h}}(\boldsymbol{a})] \approx \mathbb{E}C_{\boldsymbol{w},\boldsymbol{h}}(\mathbb{E}[\mathbf{a}]) + \mathbb{E}C'_{\boldsymbol{w},\boldsymbol{h}}(\mathbb{E}[\mathbf{a}])(\mathbf{a} - \mathbb{E}[\mathbf{a}]), \tag{31}$$

where the second term could be cancelled as $C'_{\boldsymbol{w},\boldsymbol{h}}$ is deterministic. Thus, the expectation is pushed to the mean $\mathbb{E}[\mathbf{a}] = \boldsymbol{\mu}$. The posterior form in (9) is obtained.

F Related works on uncertainty in neural network

The mutual information based uncertainty decomposition (4) is suggested by Houlsby et al. (2011). This tractable view of uncertainty is broadly used in active learning, uncertainty-critic computation vision tasks (Gal et al.; 2017b; Kendall & Gal, 2017). The followup works study how to extract distributional uncertainty, originated from the mismatch between training and testing data (Malinin & Gales, 2018). The fast approximation of expressive predictive uncertainty is studies in by Malinin et al. (2019); Cui et al. (2019). Note that the extraction of distributional uncertainty could be interesting extension to Bayes-MIL.

G More visualization results

More visualization results are shown in Fig. 9.

H Limitations

A potential limitation is the fixed feature extraction network - normally ResNet50, which is not extensively trained on the WSI data. When dealing with larger scale datasets, the features for positive patches and negative patches may not be separable by the MIL framework. Under such case, Bayes-MIL might provide marginal improvement. Future works will explore pretrained foundation models for better feature extraction.

As noted in Sec. B, the implementation of APCRF sacrifices performance for efficiency. Future works include implementation of convolution for irregular feature maps for both efficiency and performance.

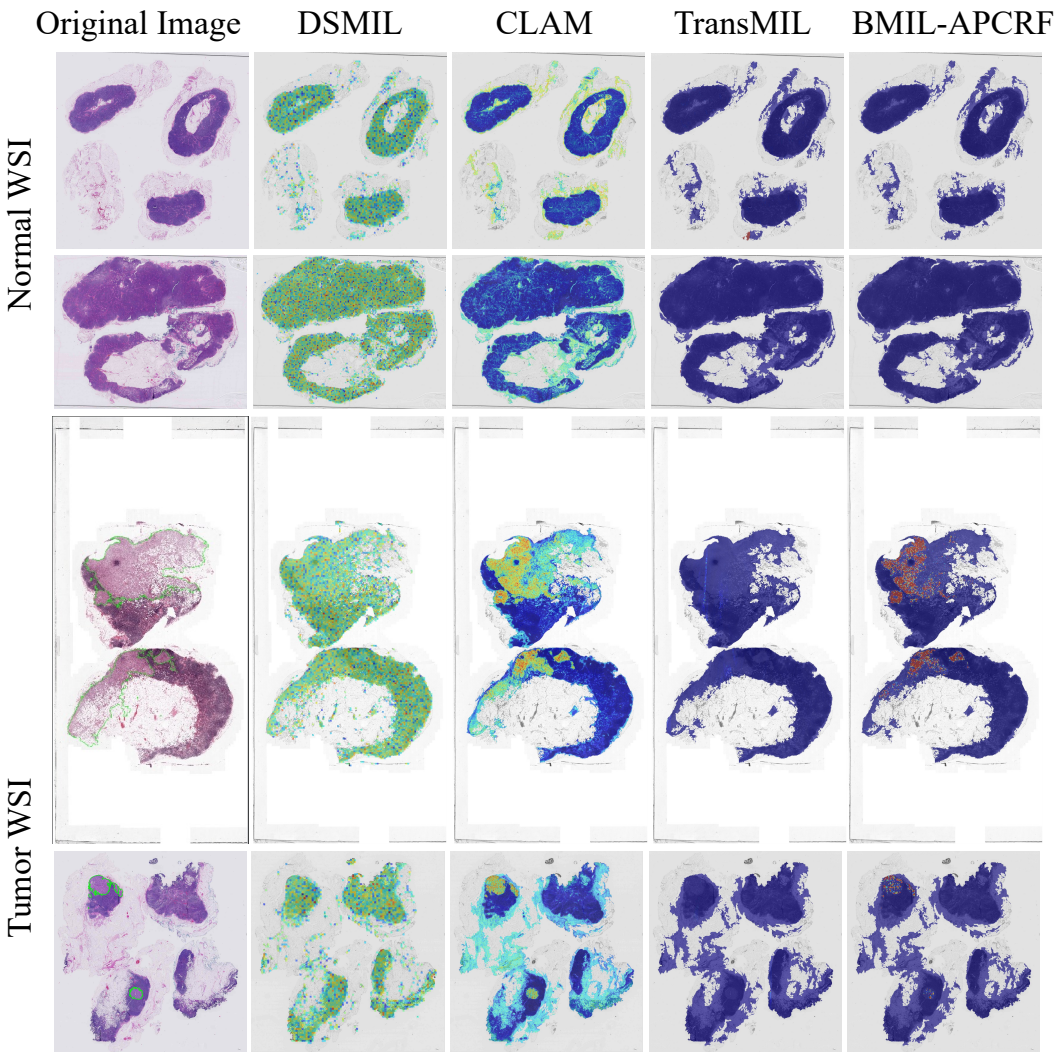


Figure 9: The visualization of normal and tumor slides and the ROIs provided by different models. The patch-level annotations for the tumor image are shown in green color in (F). The attention values $\boldsymbol{a}$ are normalized to the same range by $\frac{\boldsymbol{a}-\min_{\boldsymbol{a}}}{\max_{\boldsymbol{a}}-\min_{\boldsymbol{a}}}$. The $\min_{\boldsymbol{a}}$ and $\max_{\boldsymbol{a}}$ are the same for all methods for better visualization.