Causal ATTention Multiple Instance Learning for Whole Slide Image Classification

Xiaochun Wu WUXCH8@MAIL2.SYSU.EDU.CN
Haitao Wang WANGHT39@MAIL2.SYSU.EDU.CN
Hejun Wu* WUHEJUN@MAIL.SYSU.EDU.CN
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

We propose a new multiple instance learning (MIL) method called **C**ausal **ATT**ention **M**ultiple **I**nstance **L**earning (CATTMIL) to alleviate the dataset bias for more accurate classification of whole slide images (WSIs). There are different kinds of dataset bias due to confounders that are rooted in data generation and/or pre-training dataset of MIL. Confounders might mislead MIL models to learn spurious correlations between instances and bag label. Such spurious correlations, in turn, impede the generalization ability of models and hurt the final performance. To fight against the negative impacts of confounders, CATTMIL exploits the causal intervention using the front-door adjustment with a **C**ausal **ATT**ention (CATT) mechanism. This enables CATTMIL to remove the spurious correlations so as to estimate the causal effect of instances on the bag label. Unlike previous deconfounded MIL methods, our CATTMIL does not need to approximate confounder values. Therefore, CATTMIL is able to bring further performance boosting to existing schemes and achieve the state-of-the-art in WSI classification. Extensive experiments on classification of the two widely-used datasets of TCGA-NSCLC and CAMELYON16 show CATTMIL's effectiveness in suppressing the dataset bias and enhancing the generalization capability as well.

Keywords: Causal intervention, front-door adjustment, multiple instance learning, whole slide image classification

1. Introduction

Multiple instance learning (MIL), a kind of weakly supervised learning method, has shown its effectiveness in the computer-aided analysis of whole slide images (WSIs) Li et al. (2021); Shao et al. (2021). The analysis of WSIs plays an important role in healthcare research nowadays Mahmood et al. (2020). A WSI is generated from a WSI scanner, which projects a tissue on a biopsy slide into a gigapixel image while preserving the information of the original tissue structure He et al. (2012); Shao et al. (2021). As the size of a WSI ranges from 100 million to 10 billion pixels, much larger than a natural image Qu et al. (2023), it is infeasible to directly apply models of deep learning trained from natural images to WSIs. Therefore, the current prevailing pre-processing paradigm for WSIs is to divide a WSI into thousands of non-overlapping small patches. After pre-processing, the slide-level label is kept but the patch-level label of each patch is unavailable in most cases. Subsequently,

* Corresponding author

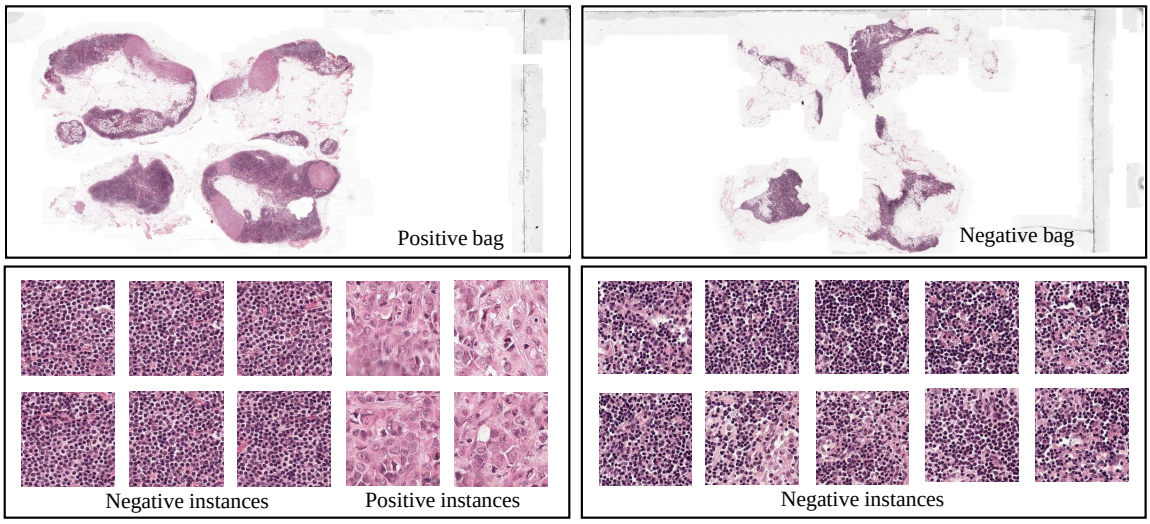
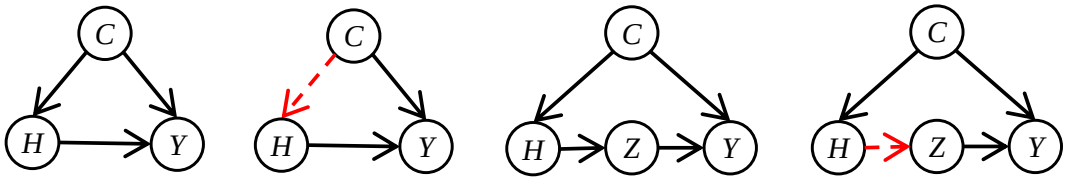


Figure 1: Illustration of differences in data generation. In patches of WSIs, positive instances are stained with the pink color, whereas negative instances appear to be the purple color. In MIL, a bag is positive if there exists at least one positive instance and negative if all instances in the bag are negative.

WSI-related tasks, such as classification and survival analysis, can be formulated as the MIL problem.

Specifically, a WSI is regarded as a bag, and the patches of the WSI are treated as instances in MIL. In general, MIL methods first perform feature extraction for patches as the instance features, and then aggregate the features for the WSI as the bag-level prediction. Among different kinds of MIL methods, the most recent are global attention-based MIL Li et al. (2021); Shao et al. (2021); Xiang and Zhang (2023). Such MIL methods have demonstrated prominent performance improvement in most WSI-related tasks. They adopt a global attention-based network as the feature aggregator that allows these MIL methods to identify cross-instance correlations so as to exploit more useful information to learn.

There is a critical problem that needs to be paid attention to in MIL methods for WSI-related tasks: dataset bias. Specifically, there are two types of factors, called confounders Pearl et al. (2016) from the causal perspective, collectively contributing to this dataset bias problem. These two classes of confounders are rooted in data generation and/or pre-training dataset, depicted as follows. (1) Confounders in data generation: Due to the differences in data generation, in patches of WSIs, positive instances are stained with the pink color, whereas negative instances appear to be the purple color, as shown in Fig. 1. The prevalence of pink positive instances and purple negative instances in the training set makes an MIL classification model exploit such co-occurrence to predict the bag label. When given



(a) Causal graph (b) $P(Y|do(H))$ (c) Causal graph (d) $P(Y|do(H))$

Figure 2: The causal graph for MIL. H: bag's instance feature embeddings, Y: bag label, C: confounders, Z: mediator. (a) The causal graph without the mediator. (b) The causal graph after back-door adjustment by cutting off the directed edge $H \leftarrow C$. (c) The causal graph with the mediator. (d) The causal graph after front-door adjustment by cutting off the directed edge $H \rightarrow Z$.

a positive test bag containing more purple instances than pink instances, the model may tend to infer that the bag is negative. Such a prediction is wrong since the model correlates color to the bag label. (2) Confounders in pre-training dataset: Most MIL models use fixed instance feature embeddings extracted by a pretrained network Li et al. (2021); Shao et al. (2021). The pretrained network is kept after pre-training while the dataset for pre-training is discarded. Studies Yang et al. (2023) have shown that this process introduces the dataset bias from the pre-training dataset into MIL and the bias cannot be estimated any more since the dataset has been dropped. In MIL, confounders mislead models to learn the spurious correlations between instances and the bag label, which is apparently harmful to the final prediction. To alleviate the negative impacts of the dataset bias in MIL, there have been a series of MIL methods being proposed Lin et al. (2023); Zhang et al. (2020). These methods greatly benefit from causal intervention Pearl et al. (2016) and causal effect.

Nevertheless, previous works need an extra training stage to approximate the values of confounders or identify stable instances. Even worse, few MIL methods consider the bias introduced by confounders in pre-training dataset. In fact, as confounders cannot be enumerated and maybe most of them are unknown, it is difficult or impossible to accurately approximate the values of confounders Liu et al. (2023).

Therefore, it is extremely challenging to achieve unbiased classification. In this paper, we propose a new bag-level MIL method called **C**ausal **ATT**ention **M**ultiple **I**nstance **L**earning (CATTMIL) to address the aforementioned challenges. As shown in Fig. 2(c), we formulate the causal relations among the confounders C, instances H, and the bag label Y into a causal graph. Inserting a mediator Z between H and Y, we set up a front-door path $H \rightarrow Z \rightarrow Y$ in the causal graph for the bag-level MIL. The front-door path provides a feasible way to exploit the causal intervention based on the front-door adjustment Pearl et al. (2016). With the front-door adjustment, CATTMIL can alleviate the confounding effect caused by the

confounders and estimate the causal effect of instances on the bag label without the need to approximate the value of the two types of confounders.

We evaluate CATTMIL on classification of two public WSI datasets, CAMELYON16 and LUSC-NSCLC. Our further ablation studies demonstrate the significant effectiveness of interventional training. The main contributions of our work are summarized as follows:

- A causal graph is proposed for the analysis of causal relations. Correspondingly, the front-door adjustment causal intervention is conducted to remove spurious correlations introduced by the confounders.
- With the front-door adjustment, CATTMIL is able to suppress the dataset bias without the need of an extra traning stage to approximate the value of confounders, whereas previous schemes do need approximation of confounders.
- CATTMIL can be employed as an add-on to general global attention-based MIL methods by only attaching an attention branch to the global attention-based MIL model. In this paper, CATTMIL has been instantiated on two baseline models. Extensive experiments have been conducted with CATTMIL, showing its effectiveness in suppressing confounders and the distinct performance improvement to baselines.

2. Related Work

2.1. Multiple Instance Learning Methods

MIL methods can be generally divided into two categories: instance-level and bag-level MIL methods. Instance-level MIL methods Hou et al. (2016); Lerousseau et al. (2020) obtain the final bag prediction through either the maximum pooling or the mean pooling of the instance probabilities. Bag-level MIL methods Ilse et al. (2018); Lu et al. (2020); Li et al. (2021); Shao et al. (2021); Xiang and Zhang (2023) aggregate instance feature embeddings in a bag into a bag feature embedding and then train a classifier upon the bag feature embedding for the bag-level prediction. Despite the simplicity, instance-level MIL methods have been demonstrated to exhibit inferior performance compared with bag-level MIL methods Zhang et al. (2022).

Recently, attention mechanisms have been the mainstream of bag-level MIL methods, where the bag feature embedding is obtained by summing the instance feature embeddings using attention scores as weights. Methods of this kind differ in the ways to generate attention scores and can be generally divided into two categories. The first category is local attention-based methods, in which ABMIL Ilse et al. (2018) utilizes an attention-based aggregation operator to assign the attention score to each instance embedding and CLAM Lu et al. (2020) further improves ABMIL with a clustering task by pulling instances with the highest and the lowest attention scores apart. Local attention-based MIL methods have obtained significant improvements and robustness. However, they typically treat instances in the bag as independent and do not consider the correlations between instances. In contrast, the second category comprehensively considers the correlations among different instances within the same bag and it is usually referred to as global attention-based methods. The following are representative example methods. DSMIL Li et al. (2021) adopts non-local pooling and obtains attention scores based on the cosine distance between each instance

to the critical instance. TransMIL Shao et al. (2021) utilizes a transformer encoder with Nyström Attention Xiong et al. (2021) to encode the mutual correlations between instances. ILRA-MIL Xiang and Zhang (2023) uses an iterative low-rank attention feature aggregator to encode cross-instance correlations. Our goal is to improve the global attention-based MIL methods with CATTMIL.

2.2. Causal Inference

Causal inference Pearl et al. (2016) has attracted increasing attention in various computer vision tasks, such as long-tailed classification Tang et al. (2020); Zhu et al. (2022), vision-language tasks Abbasnejad et al. (2020); Niu et al. (2021) and so on, for its strong ability to learn the causality between cause and effect. As for MIL classification, StableMIL Zhang et al. (2020) identifies stable instances and adds an instance to a bag as a treatment to address distribution change. IBMIL Lin et al. (2023) assumes that the confounder Pearl et al. (2016) are observable and then suppresses the bias by the back-door adjustment Pearl et al. (2016). Our CATTMIL uses the front-door adjustment Pearl et al. (2016) to remove spurious correlations and alleviate the dataset bias.

3. Method

The objective of CATTMIL is to mitigate the dataset bias for WSI classification. Our model is built upon the global attention-based MIL model and the focus of our model is to build a **C**ausal **ATT**ention (CATT) mechanism in the aggregator. Fig. 3 presents the overall framework of the proposed model, consisting of two stages. Stage 1 is the process of instance feature embedding extraction and global dictionary construction. Stage 2 trains the aggregator and classifier interventionally. In the following subsections, we give a problem formulation and then introduce the details of our method from the causal lens.

3.1. Problem Formulation

We take MIL for the binary classification as an example. Suppose $D = \{X_i\}_{i=1}^{N}$ is a dataset containing N bags, and each bag $X_i = \{x_{i,j}\}_{j=1}^{n_i}$ contains n_i instances, where each instance $x_{i,j} \in \mathbb{R}^{h\times w\times d}$ is a patch of size $h \times w \times d$. Let $\{y_{i,j}\}_{j=1}^{n_i}$ be the instance labels, the bag label Y_i is given by:

$$Y_i = \begin{cases} 0, \text{ iff} \sum_j y_{i,j} = 0 \\ 1, \text{ otherwise} \end{cases} \tag{1}$$

where $y_{i,j} \in \{0, 1\}$, $Y_i \in \{0, 1\}$. The bag label is positive if there exists at least one positive instance and negative if all instances in the bag are negative. In weakly supervised MIL, the instance labels in positive bags in the training set are unknown. Specially, for multi-class MIL classification, $Y_i \in \{0, 1, \cdots, m\}$, where m denotes the number of cancer subtypes. MIL aims to predict the bag label accurately.

Bag-level MIL methods typically comprise three modules: a feature extractor $f(\cdot)$, a feature aggregator $\sigma(\cdot)$, and a classifier $\varphi(\cdot)$. A general procedure for bag-level MIL methods

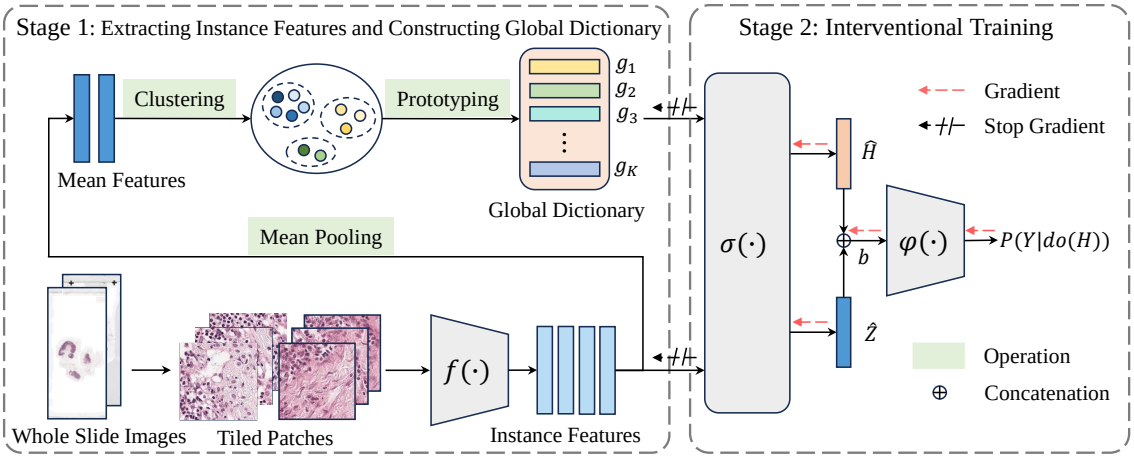


Figure 3: An overview of our proposed **C**ausal **ATT**ention **M**ultiple **I**nstance **L**earning (CATTMIL). In stage 1: First, the instance feature embeddings are extracted by a pretrained feature extractor $f(\cdot)$. Then, the global dictionary is constructed by applying a mean pooling operation on the instance feature embeddings to obtain the mean of feature embeddings and running K-means to partition these feature embeddings into clusters. In stage 2: The instance feature embeddings from the same bag and global feature embeddings from the global dictionary are fed to the aggregator $\sigma(\cdot)$ with the CATT mechanism. Then, the two output vectors of the aggregator are concatenated to obtain the final bag feature embedding, which is taken as the input for the classification by the MLP classifier $\varphi(\cdot)$.

consists of three steps: (1) The feature extractor projects each instance $x_{i,j}$ into an instance feature embedding $h_{i,j} = f(x_{i,j})$, where $h_{i,j} \in \mathbb{R}^{1\times d_1}$ and d_1 denotes the dimension of the instance feature embedding. ImageNet-pretrained ResNet He et al. (2016) and self-supervised learning-based models pretrained on histopathological images are commonly used as the feature extractors Ciga et al. (2020); Li et al. (2021); Shao et al. (2021). $H_i = \{h_{i,j}\}_{j=1}^{n_i}$ denotes a bag of instance feature embeddings. (2) The aggregator aggregates the instance feature embeddings from the same bag H_i to obtain a bag feature embedding $b_i = \sigma(H_i)$, where $b_i \in \mathbb{R}^{1\times d_2}$ and d_2 is the dimension of the bag feature embedding. (3) The Multilayer Perceptron (MLP) head, acting as the classifier, utilizes the bag feature embedding b_i to predict the bag label $\hat{Y}_i = \varphi(b_i) \in \{0, 1\}$. For clarity, i is ignored in Fig. 3.

3.2. Causal Graph for Multiple Instance Learning

To study the influence of confounders on the bag-level prediction and determine the estimation of the causal effect of instances on the bag label, we construct a causal graph for MIL to analyze the causal relations among instances, bag label, and confounders, as shown in Fig. 2(a). The causal graph contains three nodes: H: bag's multiple instance feature

embeddings, Y: bag label, and C: confounders. The directed edge represents a causal relation between two nodes: cause $\rightarrow$ effect.

$H \rightarrow Y$: The aggregation of the bag's instance feature embeddings infers the bag label.

$C \rightarrow H$: The confounders rooted in data generation and pre-training dataset have a causal effect on the bag's instance feature embeddings.

$C \rightarrow Y$: The bag label is affected by the confounders. This is because MIL is trained with instance feature embeddings-bag label pairs, which maps the information inherited from confounders into the bag label inevitably.

In the causal graph, the confounders C cause a spurious correlation between H and Y via the back-door path $H \leftarrow C \rightarrow Y$. The back-door path generates the bias when using the conventional training objective $P(Y|H)$. The back-door adjustment offers a tool for us to eliminate the spurious correlation by cutting off the directed edge $H \leftarrow C$ and calculating the likelihood $P(Y|H, c)$ within each stratum c of C if C is observable, where c denotes the value of the confounders C. The interventional training objective $P(Y|do(H))$ can be calculated as:

$$P(Y|do(H)) = \sum_{c} P(Y|H, C = c)P(C = c). \tag{2}$$

However, confounders in pre-training dataset are not accessible after pre-training. Besides, confounders in data generation are extremely complex and difficult to identify. Therefore, we can not apply the back-door adjustment to MIL directly. To address this problem, we adopt the front-door adjustment to calculate $P(Y|do(H)$ even when confounders C are unobservable. In the following, we will explain why the front-door adjustment work and show how to implement the front-door adjustment.

3.3. Causal Intervention via Front-door Adjustment

To deploy front-door adjustment, an additional mediator Z is inserted between H and Y to construct a front-door path $H \rightarrow Z \rightarrow Y$ in the causal graph for MIL, as shown in Fig. 2(c). A global attention-based model selects important information from the instance feature embeddings H to predict the bag label Y:

$$P(Y|H) = \sum_{z} P(Z = z|H)P(Y|Z = z), \tag{3}$$

where z denotes the selected knowledge from mediator Z.

To deconfound $H \rightarrow Z \rightarrow Y$, we first calculate two partial effects $P(Z|do(H))$ and $P(Y|do(Z))$. Then we chain them together to get the causal effect of H on Y:

$$P(Y|do(H)) = \sum_{z} P(Z = z|do(H))P(Y|do(Z = z)). \tag{4}$$

For $P(Z = z|do(H))$, the back-door path $H \leftarrow C \rightarrow Y \leftarrow Z$ between H and Z is naturally blocked due to the collider $C \rightarrow Y \leftarrow Z$, then we have:

$$P(Z = z|do(H)) = P(Z = z|H). \tag{5}$$

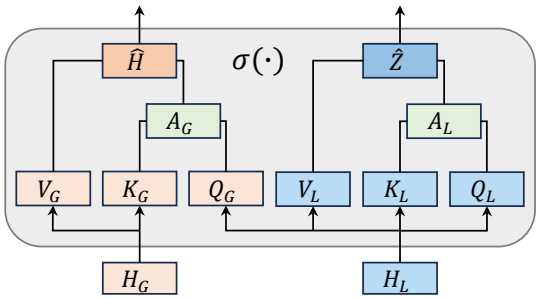


Figure 4: The structure of the causal attention module, which estimates $\hat{H}$ and $\hat{Z}$ to implement the front-door adjustment.

For $P(Y|do(Z=z))$, the back-door path $Z \leftarrow H \leftarrow C \rightarrow Y$ can be blocked by controlling H or C due to the chain $Z \leftarrow H \leftarrow C$ or the confounding $H \leftarrow C \rightarrow Y$. Since C is unavailable, we have to control H to cut off $Z \leftarrow H$ and block the back-door path, then we have:

$$P(Y|do(Z=z)) = \sum_i P(H=H_i)P(Y|H=H_i, Z=z). \tag{6}$$

After chaining two partial effects together, the intervention distribution $P(Y|do(H))$ can be calculated as:

$$\begin{aligned} &P(Y|do(H)) \\ =&\sum_z P(Z=z|do(H))P(Y|do(Z=z)) \\ =&\sum_z P(Z=z|H)\sum_i P(H=H_i)P(Y|H=H_i, Z=z). \end{aligned} \tag{7}$$

To implement front-door adjustment causal intervention into the bag-level MIL framework, we parameterize $P(Y|H,Z)$ as a network $\varphi(\cdot)$ followed by a Softmax layer that implements $P(Y|H,Z)$ as:

$$P(Y|H,Z) = \text{Softmax}(\varphi(H,Z)). \tag{8}$$

However, it is costly to compute $P(Y|do(H))$, as a tremendous large number of samples representing H and Z need to be sampled and fed into the network. To solve this problem, we employ Normalized Weighted Geometric Mean (NWGM) Xu et al. (2015) to absorb the outer sampling into the network. As a result, it needs to perform network forwarding operation only once to get the estimation $P(Y|do(H))$:

$$P(Y|do(H)) \approx \text{Softmax}(\varphi(\hat{H}, \hat{Z})), \tag{9}$$

Dataset	CAMELYON16	LUSC-NSCLC	
Subtype	/	LUSC	LUAD
No. WSIs	399	1046	
No. training WSIs	270	836	
No. test WSIs	129	210	
No. patches per bag	11,500	5,000	

Table 1: Dataset statistics for classification.

in which,

$$\begin{aligned} \hat{H} &= \sum_i P(H = H_i|\phi_1(H))H_i, \\ \hat{Z} &= \sum_z P(Z = z|\phi_2(H))z, \end{aligned} \tag{10}$$

where $\hat{H}$ and $\hat{Z}$ denote the estimation of H and Z respectively. $\phi_1(\cdot)$ and $\phi_2(\cdot)$ denote the network mapping functions.

Both $\hat{H}$ and $\hat{Z}$ can be calculated by an attention network, in which $\phi_1(\cdot)$ and $\phi_2(\cdot)$ denote query embedding functions. The attention network, jointly estimating $\hat{H}$ and $\hat{Z}$, is called **C**ausal **ATT**ention (CATT). The structure of CATT is shown in Fig. 4. We take the computation of $\hat{H}$ as an example. Since it is time-prohibitive to sample all the possible bags' instance feature embeddings to compute $\hat{H}$, we first apply the mean pooling operation on each bag's instance feature embeddings in the training set to obtain a feature embedding. Then we use K-means over all the feature embeddings, partitioning the feature embeddings into clusters. The centroid of each cluster represents a global feature embedding g_k, where $g_k \in \mathbb{R}^{1\times d_1}$ and d_1 is the dimension of the global feature embedding. Finally, we obtain a global dictionary with K global feature embeddings in the shape of $K \times d$. Let H_L represent the local instance feature embeddings that come from the current input bag and H_G represent the global feature embeddings that come from the global dictionary. The attention branch estimating $\hat{H}$ takes H_L and H_G as inputs and conditions global feature embeddings H_G on the local feature embeddings H_L. The attention branch can be expressed by the Q-K-V operation:

$$\begin{aligned} &\text{Input} : Q = H_L, K = H_G, V = H_G, \\ &\text{Prob} : A_G = \text{Softmax}((QW_Q)(KW_K)^T), \\ &\text{Output} : \hat{H} = A_G(VW_V), \end{aligned} \tag{11}$$

where W_Q, W_K, $W_V \in \mathbb{R}^{d_1\times d_1}$ are trainable parameter matrices and each attention vector a_G in A_G approximates the probability $P(H = H_i|\phi_1(H))$. Similarly, $\hat{Z}$ can be estimated as $\hat{H}$ by setting $Q = K = V = H_L$ in the other attention branch. Attention vector a_L in A_L estimates the probability $P(Z = z|\phi_2(H))$. After that, $\hat{H}$ and $\hat{Z}$ are concatenated to estimate $P(Y|do(H))$.

4. Experimental Results on Classification

We designed our experiments to answer the following questions:

- Can CATTMIL improve global attention-based MIL methods (see Fig. 5(c)), such as DSMIL Li et al. (2021) and TransMIL Shao et al. (2021), with different pretrained feature extractors (see Fig. 5(b)) for both balanced and unbalanced datasets (see Table 2 and Fig. 5(a))?
- What is the proper size of the global dictionary (see Fig. 5(d))?
- How does the implementation of the front-door adjustment affect the performance (see Table 3)?
- Can sharing parameters between the two attention branches be useful for CATTMIL (see Table 3)?
- Does CATTMIL work in a greener and more efficient way than IBMIL Lin et al. (2023) (see Table 4.6)?

4.1. Datasets

To cover balanced and unbalanced datasets, we choose TCGA-NSCLC and CAMELYON16 as datasets. Some statistical facts of chosen datasets are listed in Table 1. TCGA-NSCLC includes two subtype projects, i.e., Lung Squamous Cell Carcinoma (LUSC) and Lung Adenocarcinoma (LUAD). The dataset is divided into 836 training slides and 210 test slides following DSMIL. We directly use the patches released by DSMIL following IBMIL. There are roughly 5.2 million patches at $\times 20$ magnification, with an average of about 5,000 patches per bag. CAMELYON16 is a dataset proposed for metastasis detection in breast cancer, including 270 training slides and 129 test slides. After pre-processing, a total of about 4.6 million patches at $\times 20$ magnification, about 11,500 patches per bag are obtained.

4.2. Baselines and Evaluation Metrics

The CATTMIL can be applied to global attention-based MIL methods. Here we apply CATTMIL and IBMIL to two baselines that perform best so far: DSMIL and TransMIL. Then we compare CATTMIL with baselines w/o causal intervention and baselines w/ IBMIL. To assess the performance of the models in classification on WSI, we use accuracy and area under the curve (AUC) as the evaluation metrics. As CATTMIL is employed as an add-on to global attention-based bag-level MIL methods and the patch-level label of each patch is unavailable in most cases, patch-level performance is neither the focus nor the evaluation metric.

4.3. Implementation Details

To validate the universality of the proposed method and to compare with IBMIL fairly, we select ResNet-18 He et al. (2016), CTransPath Wang et al. (2022), and ViT-small Dosovitskiy et al. (2021) as the backbone network of the feature extractor. The ResNet-18 is pretrained on ImageNet, the CTransPath is pretrained with semantically-relevant contrastive learning (SRCL), and the ViT-small is pretrained with MoCo V3 Chen et al. (2021). The dimension of the instance feature embeddings is 512, 768, and 384 respectively. For the LUSC-NSCLC dataset, our proposed models are optimized for 50 epochs with a batch size

Method			TCGA-NSCLC		CAMELYON16	
Performance			Accuracy	AUC	Accuracy	AUC
ResNet-18 ImageNet pretrained	DSMIL		77.62	86.88	77.52	73.75
		+IBMIL	80.00	87.19	76.74	72.98
		+CATTMIL	**84.92** (±0.72)	**91.38**(±1.08)	**77.78**(±1.61)	**75.99**(±0.46)
	TransMIL		85.24	90.70	75.97	72.83
		+IBMIL	85.24	92.54	**79.07**	**79.44**
		+CATTMIL	**91.90**(±0.83)	**96.40**(±0.23)	76.36(±0.54)	74.29(±0.36)
CTransPath SRCL	DSMIL		90.95	97.13	89.15	93.26
		+IBMIL	91.43	97.51	91.47	95.20
		+CATTMIL	**92.38**(±0.48)	**97.90**(±0.18)	**93.80**(±0.78)	**96.27**(±0.50)
	TransMIL		91.90	95.55	94.57	95.88
		+IBMIL	93.81	97.24	**96.12**	97.00
		+CATTMIL	**94.92**(±0.55)	**97.98**(±0.28)	94.31(±0.44)	**97.81**(±0.29)
ViT MoCo V3	DSMIL		90.00	95.40	81.40	82.27
		+IBMIL	90.48	96.20	82.17	83.77
		+CATTMIL	**94.53**(±0.33)	**98.06**(±0.04)	**90.95**(±0.89)	**92.25**(±1.89)
	TransMIL		93.81	96.67	93.80	94.38
		+IBMIL	94.29	97.98	93.80	95.20
		+CATTMIL	**94.60**(±0.55)	**98.04**(±0.07)	**94.77**(±1.66)	**97.26**(±0.77)

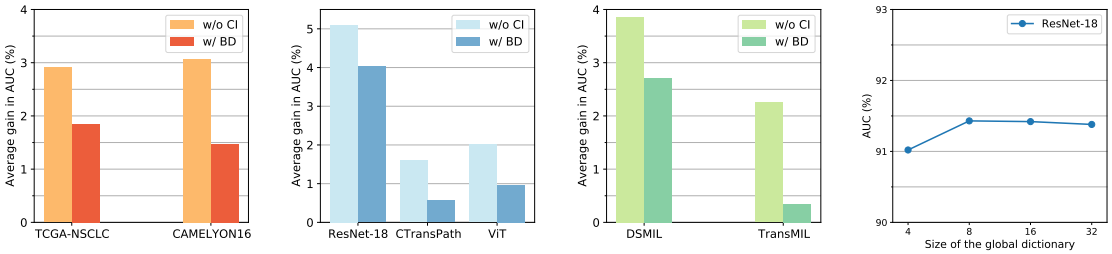
Table 2: Main results (%) on TCGA-NSCLC and CAMELYON16. The best evaluation performance is shown in boldface.

of 1. The Adam optimizer in DSMIL+CATTMIL has parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1e-4$. The Lookahead optimizer in TransMIL+CATTMIL is employed with a learning rate of 5e-5. The size of the global dictionary is set as 32. For the CAMELYON16 dataset, our proposed models are optimized for 50 epochs with a batch size of 1 and a learning rate of 1e-4, except that DSMIL+CATTMIL is trained for 200 epochs to converge with ViT-small as the feature extractor. The Adam optimizer in DSMIL+CATTMIL has parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 1e-8$. The Lookahead optimizer in TransMIL+CATTMIL is employed with a weight decay of 1e-8 and a learning rate of 1e-4. The size of the global dictionary is set as 16. Other settings are followed with their official codes. We retrain CATTMIL five times with different random seeds to report the mean and standard derivation of the evaluation performance on the test set. We perform all the experiments using PyTorch 1.13.1 on Ubuntu 20.04 with a single NVIDIA GeForce GTX 4090 GPU with 24 GB memory.

4.4. Comparison with Baselines

The classification results using different feature extractors on CATTMIL, baselines w/ IBMIL, and baselines w/o causal intervention are reported in Table 2, where the best evaluation performance is shown in boldface.

As shown in Table 2, compared with baselines w/o causal intervention, CATTMIL obtains about 3.02% improvement in accuracy and 2.99% improvement in AUC. It proves



(a) (b) (c) (d)

Figure 5: The average gain in AUC (%) (a) on datasets, (b) with pretrained feature extractors on TCGA-NSCLC, (c) on global attention-based MIL methods. "w/o CI" denotes "compared with baselines without causal intervention". "w/ BD" denotes "compared with baselines with the back-door adjustment". (d) Ablation study of the size of the global dictionary.

that CATTMIL can effectively mitigate the dataset bias in the inference stage. On the other hand, compared with baselines w/ IBMIL, CATTMIL has better performances with 2.05% improvement in accuracy and 1.64% improvement in AUC. It indicates that the front-door adjustment of CATTMIL is more effective than the back-door adjustment of IBMIL. Such observation is consistent with the discussion in section 3.2: When the confounders are unobservable, the front-door adjustment is a better choice.

In TCGA-NSCLC, positive slides contain large portions of the tumor over the whole tissue region (averagely >80% per slide). Among all global attention-based MIL methods with different feature extractors, CATTMIL achieves leading performance. As shown in Fig. 5(a), the average gain of AUC is 2.96% and 1.85% compared with baselines w/o causal intervention and baselines w/ the back-door adjustment respectively. What is more, for every feature extractor, CATTMIL improves the average performance compared with baselines w/ IBMIL, as shown in Fig. 5(b). Due to that TCGA-NSCLC is a relatively balanced dataset, the improvement is mainly the result of suppressing the bias introduced by the confounders in pre-training dataset. In particular, CATTMIL significantly improves the performance of models with ResNet-18 as the feature extractor. The reason is that ResNet-18 is pretrained on natural images, which is different from histopathological images, while CTransPath and ViT-small are self-supervised pretrained on histopathological images, which have alleviated the bias to some degree during pre-training. In CAMELYON16, each positive slide contains only a small portion of the tumor (averagely <10% per slide), thus there is a highly imbalanced distribution of positive and negative instances in a positive bag. Despite imbalanced distribution, CATTMIL still obtains leading performance in 5 of 6 cases. In detail, CATTMIL achieves 3.07% higher in AUC than baselines w/o causal intervention and 1.47% higher in AUC than baselines w/ the back-door adjustment, as shown in Fig. 5(a). The above observations demonstrate that causal intervention via the front-door adjustment indeed mitigates the bias introduced by confounders in data generation and pre-training dataset without the need to approximate the confounders.

Setting	Precision	Recall	Accuracy	AUC
baseline	81.98	86.25	80.00	87.19
K=4	86.37	85.52	84.44	91.02
K=8	86.17	**86.74**	85.24	**91.43**
K=16	86.44	86.60	**85.71**	91.42
K=32	**87.35**	85.29	84.92	91.38
$\oplus$	**87.35**	85.29	**84.92**	**91.38**
$+$	84.27	83.90	81.91	89.75
-	86.24	**85.63**	84.44	90.37
w/o sharing	**87.35**	85.29	**84.92**	**91.38**
w/ sharing	83.97	**86.09**	83.18	90.00

Table 3: Results (%) of ablation on model design variants. The best evaluation performance is shown in boldface.

As shown in Fig. 5(c), both DSMIL and TransMIL w/ the front-door adjustment outperform baselines on average. It indicates that CATTMIL is universal to global attention-based MIL methods and consistently enhances their performance. Besides, the improvement of DSMIL is greater than that of TransMIL. This is probably due to that the structure of DSMIL is simpler than TransMIL, making DSMIL easier to suffer from the negative impacts of the confounders.

Overall, CATTMIL consistently improves all global attention-based MIL methods with all feature extractors on both datasets, which demonstrates the effectiveness and universality of CATTMIL.

4.5. Ablation on Model Design Variants

In this section, we ablate the important design elements in the proposed method. Experiments are conducted on the TCGA-NSCLC dataset with ResNet-18 as the feature extractor and DSMIL as the backbone of the aggregator.

Size of the global dictionary. We study the effect of different sizes of the global dictionary. The results are shown in Fig. 5(d) and Table 3. We observe that the performance of our method is quite stable with the size of the global dictionary K varying from 4 to 32, which indicates that we do not need to elaborately tune this hyper-parameter.

Implementation of the front-door adjustment. We study how different implementations of front-door adjustment affect the effectiveness of the proposed method. Specifically, given $\hat{H}$ and $\hat{Z}$, we explore three different ways to combine them, denoted as $\hat{H} \star \hat{Z}$, where $\star \in \{\oplus, +, -\}$. $\oplus$ is concatenation. $+/-$ is element-wise addition/subtraction. The results are summarized in Table 3. We can see that all these implementations can boost the performance. Such observation underscores the effectiveness of the proposed method. Among the three combining modes, the concatenation mode performs the best. The reason is that the concatenation mode may maintain original information from $\hat{H}$ and $\hat{Z}$ better after combining them. Therefore, $\hat{H} \oplus \hat{Z}$ is fed into the MLP classifier in experiments.

Sharing parameters. We investigate the effect of sharing parameters between two attention branches. The results are shown in Table 3. Sharing the parameters makes the

Method	Time (h)	Accuracy	AUC
DSMIL	1.83	77.62	86.88
+IBMIL	3.67	80.00	87.19
+CATTMIL	1.81	84.92	91.38

Table 4: Time comparison of different methods.

performance decrease by 1.74% (from 84.92% to 83.18%) in accuracy and 1.38% (from 91.38% to 90.00%) in AUC. This observation suggests that $\hat{H}$ and $\hat{Z}$ do not need to stay in the same representation space. Therefore, the parameters of the two branches are designed to be independent in our model.

4.6. Computational Cost Analysis

To evaluate the efficiency of our method, we compare the training time of IBMIL and CATTMIL on the TCGA-NSCLC dataset. The comparison results are shown in Table 4.6. We can train the CATTMIL about 2.0x faster than IBMIL. This is because CATTMIL does not need an extra stage of retraining aggregator to approximate the confounders. These experiments, combined with the results in Table 2, verify that our approach can achieve outstanding performance with impressive training efficiency.

5. Conclusion

In this paper, we first introduce a causal graph to analyze causal relations. Then we employ the causal intervention for a comprehensive analysis of deconfounded multiple instance learning (MIL). With these analysis, we propose a method called **C**ausal **ATT**ention **M**ultiple **I**nstance **L**earning (CATTMIL) using the front-door adjustment to alleviate the confounding effect. Extensive experiments on classification have been conducted on global attention-based MIL methods. The results show that CATTMIL can remove the spurious correlations between instances and the bag label and boost the performance of global attention-based MIL methods. Our future research direction is designing a general MIL framework based on the front-door adjustment to adapt to both global attention-based and local attention-based MIL methods.

Acknowledgments

This research was funded by [GuangDong Basic and Applied Basic Research Foundation] grant number [2023A1515140146].

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *CVPR*, pages 10041–10051, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629, 2021.

Ozan Ciga, Anne L. Martel, and Tony Xu. Self supervised contrastive learning for digital histopathology. *Preprint arXiv:2011.13971*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Lei He, L. Rodney Long, Sameer K. Antani, and George R. Thoma. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine*, 107(3):538–556, 2012.

Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *CVPR*, pages 2424–2433, 2016.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, volume 80, pages 2132–2141, 2018.

Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *MICCAI*, volume 12265, pages 470–479, 2020.

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 *CVPR*, pages 14318–14328, 2021.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *CVPR*, pages 19830–19839, 2023.

Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 45(10):11624–11641, 2023.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Preprint arXiv:2004.09666*, 2020.

Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. McKay, Kevan J. Salimian, Alexander S. Baras, and Nicholas J. Durr. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11):3257–3267, 2020.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *CVPR*, pages 12700–12710, 2021.

Judea Pearl, Madelyn Glymour, Nicholas Jewell, Alex Balke, David Chickering, David Galles, Dan Geiger, Moises Goldszmidt, Jin Kim, George Rebane, Ilya Shpitser, Jin Tian, Thomas Verma, Elias Bareinboim, Bryant Chen, Andrew Forney, Ang Li, and Karthika Mohan. Causal inference in statistics a primer. 2016.

Linhao Qu, Yingfan Ma, Xiaoyuan Luo, Manning Wang, and Zhijian Song. Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. *Preprint arXiv:2307.02249*, 2023.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, pages 2136–2147, 2021.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *ICLR*, 2023.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI*, pages 14138–14148, 2021.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 37, pages 2048–2057, 2015.

Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 45(11):12996–13010, 2023.

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 *CVPR*, pages 18780–18790, 2022.

Weijia Zhang, Lin Liu, and Jiuyong Li. Robust multi-instance learning with stable instances. In *ECAI*, volume 325, pages 1682–1689, 2020.

Beier Zhu, Yulei Niu, Xian-Sheng Hua, and Hanwang Zhang. Cross-domain empirical risk minimization for unbiased long-tailed classification. In *AAAI*, pages 3589–3597, 2022.