Robust Self-Supervised Multi-Instance Learning with Structure Awareness

Yejiang Wang[1], Yuhai Zhao[1,*], Zhengkui Wang[2], Meixia Wang[1]

[1]School of Computer Science and Engineering, Northeastern University, China
[2]InfoComm Technology Cluster, Singapore Institute of Technology, Singapore
wangyejiang@stumail.neu.edu.cn, zhaoyuhai@mail.neu.edu.cn,
zhengkui.wang@singaporetech.edu.sg, wangmeixia@stumail.neu.edu.cn

Abstract

Multi-instance learning (MIL) is a supervised learning where each example is a labeled bag with many instances. The typical MIL strategies are to train an instance-level feature extractor followed by aggregating instances features as bag-level representation with labeled information. However, learning such a bag-level representation highly depends on a large number of labeled datasets, which are difficult to get in real-world scenarios. In this paper, we make the first attempt to propose a robust Self-supervised Multi-Instance LEarning architecture with Structure awareness (SMILES) that learns unsupervised bag representation. Our proposed approach is: 1) permutation invariant to the order of instances in bag; 2) structure-aware to encode the topological structures among the instances; and 3) robust against instances noise or permutation. Specifically, to yield robust MIL model without label information, we augment the multi-instance bag and train the representation encoder to maximize the agreement between the representations of the same bag in its different augmented forms. Moreover, to capture topological structures from nearby instances in bags, our framework learns optimal graph structures for the bags and these graphs are optimized together with message passing layers and ordered weighted averaging operator towards contrastive loss. Our main theorem characterizes the permutation invariance of the bag representation. Compared with state-of-the-art supervised MIL baselines, SMILES achieves average improvement of 4.9%, 4.4% in classification accuracy on 5 benchmark datasets and 20 newsgroups datasets, respectively. In addition, we show that the model is robust to the input corruption.

Introduction

Multi-instance learning (MIL) is a form of supervised learning where training instances are arranged in sets, namely bags, and each bag is assigned a binary label (Yuan et al. 2021; Pal et al. 2022; Huang et al. 2022). The standard MIL assumption is that a bag is positive if it contains at least one positive instance, and negative otherwise.

In the past decades, much research effort has been devoted to improve the performance of MIL by learning the bag-level representation, which implicitly utilizes bag-to-bag similarity or explicitly trains a bag classifier (Feng et al. 2021; Wang et al. 2018). For large-scale MIL scenarios like drug activity prediction, where each molecule can be represented as a bag and the instances correspond to different conformations (molecular structures) of that compound, these methods often implement the bag-level MIL models by a two-stage strategy: first training an instance-level feature extractor, and then aggregating features as bag-level representations with label information (Ilse, Tomczak, and Welling 2018). However, it may be difficult to collect a multi-instance learning datasets composed of fully labeled bags in real-world applications, due to the significant labeling costs. For example, high-quality molecule data with human labeling could be costly and it is difficult, if not impossible to create fully labeled datasets with millions of molecules (Rong et al. 2020; Zhang et al. 2021). To tackle this challenge, in this paper, we make an attempt to learn the representation of bag in a self-supervised manner without the requirement of any label information.

A robust self-supervised MIL learning for bag representation should fulfil below important properties. First, it should generate the bag representation that is invariant to the permutation of the set of instances. The example (i.e. a bag) in MIL is described by a set of feature instances. The order independence of set can be used to design models with improved efficiency and generalization (Wagstaff et al. 2022; Maron et al. 2019). Second, it should have the capability of capturing the topological structure information on tasks where the objects have inherent interactions. Previous studies on multi-instance learning typically treated instances in the bags as independently and identically distributed (Huang et al. 2022; Feng et al. 2021; Ilse, Tomczak, and Welling 2018). However, the local structures and the proximity information from nearby instances are important in MIL model (Zhang 2021). Third, it should be able to handle the bag noise. It is inevitable that the provided multi-instance bags are incomplete and noisy in real-world scenarios (Chevaleyre and Zucker 2000; Luengo et al. 2021). Hence, developing robust multi-instance learning models to resist unnoticeable perturbation (e.g., missing or error instances in bags) is of significant importance.

In this paper, we provide a full characterization of multi-instance learning and present a robust Self-supervised Multi-Instance LEarning architecture with Structure awareness (SMILES) to capture all the above properties. Specifically,

*Corresponding author.

to yield robust MIL model without using labels, we augment the multi-instance bag and train the representation encoder to maximize the correspondence between the representations of the same bag in its different augmented forms. Through maximizing their consistency, our model is robust with respect to noise/perturbation of data. To capture the geometric structures of instances in bags, we generate learnable graph adjacency matrices for the bags, which well respects the node proximity conveyed by the original feature instances, and these graphs are optimized together with message passing layers and the ordered weighted averaging operator towards contrastive loss, where informative hidden connections can be discovered. Our main theorem characterizes the permutation invariance of the representation of node and bag. In summary, our core contributions are three-fold:

- We propose an unsupervised learning paradigm SMILES for multi-instance learning, which is more practical and challenging than the existing supervised counterpart. To the best of our knowledge, this is the first attempt to learn the bag representation in an unsupervised setting.
- Our self-supervised model provides approaches to augment the bag to offer noise robustness when the MIL data have possibly noise. Together with the graph generation for the multiple instances, the bag representation can be associated with structure awareness, while be invariant to permutations of its constituent instances.
- Extensive experiments show that SMILES, the self-supervised model significantly outperforms state-of-the-art supervised MIL models, and is robust against common injected noise/permutation.

Related Studies and Preliminaries

In this section, we briefly review related studies and introduce preliminary knowledge.

Multi-Instance Learning. In traditional supervised learning, each learning example consists of a fixed number of values (i.e., an instance) with a label. However, in many applications, only a bag of instances is given a label, which is referred as multi-instance learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez 1997). In this paper, we follow the notation of (Gärtner et al. 2002). $\mathcal{X} \subset \mathbb{R}^{d_{\text{in}}}$ denotes the instance space, Ω is the set of labels y. In MIL, the label is assumed to be binary, so $\Omega = \{\top, \bot\}$. A multi-instance concept is a map $\nu_{mi} : 2^{\mathcal{X}} \to \Omega$ defined as

$$\nu_{mi}(X) \Leftrightarrow \exists x \in X : c(x) \tag{1}$$

where $c \in \mathcal{C}$ is a concept from a concept space $\mathcal{C}$, and $X \subseteq \mathcal{X}$ is a set of instances. Supervised MIL problem aims to predict labels of new bags based on the labeled training dataset $\mathbb{D} = \{(X, y)\}$ (Lin et al. 2022; Chu et al. 2020).

In this work, we are interested in learning multi-instance with no supervision. Self-supervised learning (SSL), emerging as a learning paradigm that can enable training on massive unlabeled data, recently has received considerable attention (Von Kügelgen et al. 2021; Reed et al. 2022). However, as far as we know, there is no work explored self-supervised learning for multi-instance problem. Formally, *self-supervised multi-instance learning* should learn bag representation by a function $f_{rep} : 2^{\mathcal{X}} \to \mathbb{R}^{d_{\text{out}}}$ that transforms the multi-instance of a bag X into a d_{out}-dimensional instance space $f_{rep}(X) = (a_1, ..., a_{d_{\text{out}}})$ without label.

Multi-instance Noise. Obtaining multi-instance model that are robust to perturbation/noise has been an active topic of research (Chevaleyre and Zucker 2000; Luengo et al. 2021). In this paper, we study the set of training perturbations: $\mathfrak{U}_p(\Lambda) = \left\{\delta \in \mathbb{R}^{d_{in}} : \|\delta\|_p \leq \Lambda\right\}$, where δ denotes the measurement error, p is the ℓ_p-norm and Λ controls the amplitude of the perturbations. Formally, the generation of noise is a black-box feedback mechanism which, when called at instance x, returns a random vector $g(x; \delta)$ with δ drawn from some (complete) probability space $(\mathfrak{U}_p(\Lambda), \mathcal{F}, \mathbb{P})$, which is independent of the value of x. Therefore, the oracle draws an i.i.d. sample $\delta \in \mathfrak{U}_p(\Lambda)$ and returns an observed instance: $g(x; \delta) = x + \delta$. In supervised setting, the simplest and most straightforward way to defend against such noise is to minimize the loss of measurement error examples

$$\underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{(X,y)\sim\mathbb{D},\delta\in\mathfrak{U}_p(\Lambda)}\, \mathcal{L}_{ce}(\theta, \{g(x; \delta)|x \in X\}, y) \tag{2}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss, and θ is parameters.

Multi-instance Structure. Multi-instance structure learning targets jointly learning graph structure and corresponding representation to improving the expressiveness of MIL models (Pal et al. 2022; Zhao et al. 2021). It aims to learn functions of sets of n instances X into graph with n nodes. Let $G = (V, \mathcal{E}, X')$ be an undirected graph with respect to X, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices and suppose that every vertex v_i corresponds to a d'-dimension vector $x'_i \in X'$, and $\mathcal{E}$ is a similarity matrix of vertexes, the element $e_{i,j}$ denotes the weight of edge. Another way to see graph with features is to see the graph as tensors of order 2: $G \in (\mathbb{F}^n, \mathbb{F}^{n^2})$, where $\mathbb{F}$ denotes arbitrary finite-dimensional space of the form $\mathbb{R}^q$ (for various values of q) typically representing the feature space. Here, $X' \in \mathbb{F}^n$ and $\mathcal{E} \in \mathbb{F}^{n^2}$. *Multi-instance structure learning* considers learning a function $f_{sl} : 2^{\mathcal{X}} \to (\mathbb{F}^n, \mathbb{F}^{n^2})$ maps the input space $2^{\mathcal{X}}$ to the graph space. Intuitively, if x_i and x_j are nearest neighbors on $X \in 2^{\mathcal{X}}$ with a high degree of similarity, the corresponding vertexes should be close to one another.

Permutation Invariance. In multi-instance learning tasks the representation of bag that we want to learn should be invariant to any permutation of the instances in bag. In addition, for the learned multi-instance structure (i.e. graph), it is importance to ensure that the model remains permutation invariant to the structure (Zhang et al. 2022; Zaheer et al. 2017; Wagstaff et al. 2019).

For a multi-instance bag, an instance permutation action $\pi \in \mathbb{S}_{\mathcal{B}}$ is a left action $\phi : \mathbb{S}_{\mathcal{B}} \times 2^{\mathcal{X}} \to 2^{\mathcal{X}}$ with the element π on a sorted sequence of n instances represented as $X = (x_1, \ldots, x_n)$ of a bag to output a corresponding permuted sequence of instances i.e., $\phi(\pi, X) = (x_{\pi(1)}, \ldots, x_{\pi(n)})$. A map $f : 2^{\mathcal{X}} \to \mathbb{R}^{d_{\text{out}}}$ satisfying $f \circ \phi(\pi, X) = f(X)$ for all $\pi \in S_{\mathcal{B}}$ and $X \in 2^{\mathcal{X}}$ is called *permutation invariant*. For the graph generated from bag, a vertex permutation action $\pi \in \mathbb{S}_{\mathcal{G}}$ is defined in similar way: $\phi : \mathbb{S}_{\mathcal{G}} \times V \to V$, and $\phi(\pi, V) = (v_{\pi(1)}, \ldots, v_{\pi(n)})$. The permutation action $\pi \in \mathbb{S}_{\mathcal{G}}$ also acts on any vector defined over the nodes V, i.e.,

$(x_i) \in \mathbb{F}^n$, and output an equivalent vector with the order of the nodes permuted i.e., $(x_{\pi_i}) \in \mathbb{F}^n$. A function f acting on a graph G given by $f : (\mathbb{F}^n, \mathbb{F}^{n^2}) \to \mathbb{R}^{d_{\text{out}}}$ is *$\mathcal{G}$-invariant* whenever it is invariant to any vertex permutation action $\pi \in \mathbb{S}_{\mathcal{G}}$ in the $(\mathbb{F}^n, \mathbb{F}^{n^2})$ graph space i.e., $f \circ \phi(\pi, V) = f(V)$ and all isomorphic graphs obtain the same representation.

Methodology

In this section, we will present a robust self-supervised multi-instance learning method with structual awareness, named SMILES. Given an input bag, SMILES aims to learn the self-supervised representation of the bag through maximizing the consistency between two augmented views of the input bag via contrastive loss in the latent space. To capture the structural relations among instances we generate multi-instance graph in a learnable manner. The bag representation is obtained by encoding the graph with message passing layers and the ordered weighted averaging operator, where permutation invariant of unified representaion encoder for the bag is theoretical guaranteed. We summarize all the steps of our framework in Algorithm 1.

Bag Augmentation

A way of inducing inductive bias for multi-instance learning is data augmentation, which we use in the bag data and which plays a prominent role in robustness learning overall. Define $\mathcal{A} : 2^{\mathcal{X}} \to 2^{\mathcal{X}}$ as the function class of augmentations and $\mathcal{F}_{rep}$ as the class of representation encoders. For $f_a \in \mathcal{A}$ and any $X \in 2^{\mathcal{X}}$, we define

$$f_a(X) = \left\{ \tilde{x} \,\middle|\, \tilde{x} = g(x; \delta), x \in X, \delta \in \mathfrak{U}_p(\Lambda) \right\} \quad (3)$$

Suppose that there is $f \in \mathcal{F}_{rep}$ satisfies $f(f_a(X)) = f(X)$ for X. The noise perturbation can provide contrastive information in various magnitude for the encoder to learn the representations. Thus, the self-supervised noise-against multi-instance learning objective for an instance-wise perturbation, following the supervised formulation of Eq.(2), could be given as follows

$$\underset{\theta}{\operatorname{argmin}} \; \mathbb{E}_{(X)\sim\mathbb{D}} \; \mathcal{L}_\theta\Big(X, \{f_a(X)\}, \{X^-\}\Big) \quad (4)$$

where $\{X^-\}$ are the negative bags for X, which are bags of other examples, and the contrastive loss $\mathcal{L}_\theta$ can be defined

$$\begin{aligned} &\mathcal{L}_\theta\Big(X, \{X^+\}, \{X^-\}\Big) \\ &:= -\log \frac{\sum_{\{z^+\}} \exp\left(\cos\left(z, z^+\right)/\tau\right)}{\sum_{z^\circ \in \{z^+, z^-\}} \exp\left(\cos\left(z, z^\circ\right)/\tau\right)} \end{aligned} \quad (5)$$

where τ is a temperature, $\cos(u, v) = u^T v / \|u\| \|v\|$ denotes cosine similarity, $z, \{z^+\}$ and $\{z^-\}$ are corresponding latent vectors obtained by the representation encoder $z = f_{rep}(X)$, $\{z^+\}$ and $\{z^-\}$, respectively.

Bag Structure Awareness

The construction of a meaningful graph topology plays a crucial role in the effective representation and analysis of multi-instance data. However, a natural choice of the graph is not readily available from bag and it is thus desirable to infer or learn a graph topology from the instances in the bag.

We generate the feature similarity matrix $\mathcal{E} \in \mathbb{F}^{n^2}$ for determining the possibility of an edge between two nodes based on node features. Specifically, for each node v_i with feature vector $x_i \in X$, we adopt a non-linear feature mapping layer $f_{nl} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{F}$ to project the feature x_i to a d'-dimension latent feature

$$x'_i = f_{nl}(x_i) \coloneqq \sigma\left(x_i \cdot W_{nl} + b_{nl}\right) \quad (6)$$

where $\sigma(\cdot)$ denotes a non-linear activation function, $W_{nl} \in \mathbb{R}^{d_{in} \times d'}$ and $b_{nl} \in \mathbb{R}^{1 \times d'}$ denote the mapping matrix and the bias vector, respectively. Then, we perform metric learning on the latent features and obtain the learned feature similarity graph $\mathcal{E} \in \mathbb{F}^{n^2}$ where the edge between nodes v_i and v_j is obtained by

$$\mathcal{E}[i, j] = s\left(x'_i, x'_j\right) \times [\![s\left(x'_i, x'_j\right) \geq \epsilon]\!] \quad (7)$$

where $[\![\cdot]\!]$ is the Iverson bracket, i.e., 1 whenever a condition in the bracket is satisfied, and 0 otherwise. $\epsilon \in [0, 1]$ is the threshold that controls the sparsity of feature similarity graph, and larger ϵ implies a more sparse feature similarity graph. s is a K-head weighted cosine similarity function

$$s\left(x'_i, x'_j\right) = \frac{1}{K} \sum_{k}^{K} \cos\left(w_k \odot x'_i, w_k \odot x'_j\right) \quad (8)$$

where $\odot$ denotes the Hadamard product, and $W_{kh} = [w_k]$ is the learnable parameter matrix of s that weights the importance of different dimensions of the lantern feature vectors. By performing metric learning as in Eq.(7) and ruling out edges with little feature similarity by threshold ϵ, we learn the candidate feature similarity graph $(X', \mathcal{E}) = f_{sl}(X)$.

Bag Representation

Given a graph $G = (V, \mathcal{E}, X')$ generated above, to learn vertex representations for every vertex $v \in V$, in this work, we use a message passing framework on the generated graph, which preserves adjacency information between nodes as follows. Let $h_i^\ell \in \mathbb{F}_\ell$ denotes the feature at layer ℓ associated with node i, the updated feature $h_i^{\ell+1}$ is obtained as: $h_i^{\ell+1} = f_{upd}(h_i^\ell, \{\!\{h_j^\ell \mid j \in \mathcal{N}_i\}\!\})$, where $j \in \mathcal{N}_i$ means that nodes j and i are neighbors in the graph G, i.e. $(i, j) \in \mathcal{E}$, and the function $f_{upd} : 2^{\mathbb{F}_\ell} \to \mathbb{F}_{\ell+1}$ is a learnable function taking as input the feature vector of the center vertex h_i^ℓ and the multiset of features of the neighboring vertices $\{\!\{h_j^\ell \mid j \in \mathcal{N}_i\}\!\}$. Indeed, for any such function f_{upd} can be approximated by a layer of the form

$$h_i^{\ell+1} = \sigma\left(W^\ell \cdot \left(h_i^\ell \otimes f^\ell\left(h_i^\ell, \{\!\{h_j^\ell \mid j \in \mathcal{N}_i\}\!\}\right)\right)\right) \quad (9)$$

where $f^\ell : 2^{\mathbb{F}_\ell} \to \mathbb{F}_{\ell+1}$ is injective set funcitons in the ℓ-th layer, $\otimes$ denotes vector concatenation, W^ℓ is learnable weight matrix and σ is an element-wise activation function. We get the ℓ-th message passing layer $f_{mp}^\ell : \mathbb{F}_\ell \to \mathbb{F}_{\ell+1}$ (note that f_{mp} depends implicitly on the graph/edge). Then,

by the composition of f_{mp}^{ℓ}, we obtain the novel representation of each instance in bag

$$x_i'' = f_{mp}^{L} \circ \dots f_{mp}^{2} \circ f_{mp}^{1}(x_i') \tag{10}$$

where L denotes the total number of layers used.

Theorem 1 (Node Representation)**.** *For a node-featured graph $G = (V, X', \mathcal{E}) \in \mathbb{F}^n \times \mathbb{F}^{n^2}$ and a vertex representation function $\psi(v, V, \mathcal{E}, X') : V \times (\mathbb{F}^n, \mathbb{F}^{n^2}) \to \mathbb{R}^{d_{\text{out}}}$ on $v \in V$ given by Eq.(10). Then, for all permutation actions $\forall \pi \in \mathbb{S}_{\mathcal{G}}$,*

$$\psi(v, V, \mathcal{E}, X') = \psi(\phi(\pi, v), \phi(\pi, V), \mathcal{E}, \phi(\pi, X')).$$

This implies that the map ψ is a $\mathcal{G}$-invariant for any node.

Proof. For any two different vertex permutation actions $\pi, \pi' \in \mathbb{S}_{\mathcal{G}}$ supposed they are satisfied $\psi(\phi(\pi, v), \phi(\pi, V), \mathcal{E}, \phi(\pi, X')) \neq \psi(\phi(\pi', v), \phi(\pi', V), \mathcal{E}, \phi(\pi', X'))$. This indicates that for different order of the nodes in graph the same vertex may get different representations. For each layer f_{mp}^{ℓ} on the vertex v there is a corresponding map $f_{mp,v}^{\ell} : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_{\ell+1}}$. Let $\ell = 1$, expanding $f_{mp,v}^{\ell}$ for vertex permutation actions and applying the cancellation law of groups, since $h_v^0 = x_v'$ is identical for these two permutations, it means the h_v^1 is equivalent as well. However, according to the previous assumption this is not possible. By induction on $\ell \geq 2$, if the contradiction holds for some ℓ, it holds for $\ell + 1$ as well, which conclude our proof. □

However, the above process generates only representations of instances in bag, hence aggregation operation is required, which tries to summarize those representations into a single element. To do this, the representation of any bag are obtained using the ordered weighted averaging (OWA) operator (Yager 1988) $f_{owa} : \mathbb{F}_{L+1}^n \to \mathbb{R}^{d_{\text{out}}}$ as

$$z = f_{owa}\left(\{x_i'' \mid i \in [n]\}; \boldsymbol{\zeta}\right) = \sum_{i=1}^{n} \zeta_i x_{(i)}'' \tag{11}$$

where $x_{(i)}''$ is the i-th largest element in the set $\{x_i'' \mid i \in [n]\}$, and $\boldsymbol{\zeta} = [\zeta_1 \dots \zeta_n]$ is a parameter vector associated with f_{owa}, such that ζ_i is nonnegative and $\sum_{i=1}^{n} \zeta_i = 1$. The OWA operator can be seen as a generalization of any aggregation operation that can be made over a set of values. For example, the maximum operator over a set of values can be modeled with the weight vector $\langle 1, 0, \dots, 0 \rangle$. In this work we choose $\langle 1/n, \dots, 1/n \rangle$ for averaging aggregation. Based on the Theorem 1 and the permutation invariance of OWA operator, we can easily get following result.

Theorem 2 (Bag Representation)**.** *The representation encoder $f_{rep}(X) = f_{owa}(\{f_{mp}^{L} \circ \dots f_{mp}^{2} \circ f_{mp}^{1} \circ f_{sl} \circ g(x_i; \delta) \mid x_i \in X\})$ is permutation invariant for bag X.*

We further introduce several practical data transformations to approximate the perturbation distribution in Eq.(3).

Remark (Bag Augmentation)**.** *For the model f_{rep}, randomly dropping, masking, replacing, or randomizing instances in bags are all special cases of bag augmentation in Eq.(3), where dropping randomly removes certain ratio of instances, masking randomly set certain ratio of instance elements as zero, replacing randomly replaces certain ratio of elements of an instance with the corresponding elements of another randomly choosing instance in bag, randomizing randomly assigns random vectors to certain ratio of instances.*

Algorithm 1: **SMILES**.

input: unlabeled training data $X \subseteq \mathcal{X}$, batch size N, temperature τ, augmentation ratio c, encoder network f_{rep}, pre-train head network g.
for sampled mini-batch $\{X_i\}_{i=1}^{N} \subseteq \mathcal{X}$ **do**
  for $i \in [N]$, $\check{X}_i = f_a(X_i)$, $\tilde{X}_i = f_a(X_i)$.
  # generate corrupted views.
  let $(\check{X}_i', \check{\mathcal{E}}_i) = f_{sl}(\check{X}_i)$, $(\tilde{X}_i', \tilde{\mathcal{E}}_i) = f_{sl}(\tilde{X}_i)$, $\forall i \in [N]$.
  # generate multi-instance graph.
  let $\check{z}_i' = g(f_{rep}(\check{X}_i'))$, $\tilde{z}_i' = g(f_{rep}(\tilde{X}_i'))$, $\forall i \in [N]$.
  # embeddings for views.
  let $t_{i,j} = \check{z}_i'^{\top} \tilde{z}_i' / \left(\|\check{z}_i'\|_2 \cdot \|\tilde{z}_j'\|_2\right), \forall i, j \in [N]$.
  # pairwise similarity.
  define $\mathcal{L}_\theta := \frac{1}{N} \sum_{i=1}^{N} -\log\left(\frac{\exp(t_{i,i}/\tau)}{\frac{1}{N}\sum_{k=1}^{N} \exp(t_{i,k}/\tau)}\right)$.
  update networks f_{rep} and g to minimize $\mathcal{L}_\theta$ by SGD.
end for
return encoder network f_{rep}.

Architecture. Based on above analysis, we propose SMILES framework (Algorithm 1) as following:
(i) **Bag data augmentation.** The given bag X undergoes bag augmentations to obtain two correlated views $\check{X}, \tilde{X} := f_a(X)$, as a positive pair. In practice, according to Ramark, in our work, view augmentation methods include instance dropping, adding, randomizing, or masking.
(ii) **Bag structure awareness.** The multi-instance graph is generated by $(X', \mathcal{E}) = f_{sl}(X)$ according to Eq.(6) and (7) for the augmented bags, which capture the structural interactions between instances in the bags.
(iii) **Encoder.** A representation encoder $f_{rep}(\cdot)$ extracts bag-level representation vectors $\check{z}, \tilde{z}$ for the augmented bags using the above graphs $\check{X}', \tilde{X}'$. The message passing layers in two encoders share parameters in the pre-training.
(iv) **Projection head.** A non-linear function $g(\cdot)$ named projection head maps representations to another latent space where the contrastive loss is calculated. In our work, a two-layer perceptron is applied to obtain $\check{z}' = g(\check{z}), \tilde{z}' = g(\tilde{z})$.
(v) **Contrastive loss.** A contrastive loss function $\mathcal{L}_\theta(\cdot)$ (in Eq.(5)) is defined to enforce maximizing the consistency between positive pairs $\check{z}', \tilde{z}'$ compared with negative pairs.

Experiments

We empirically evaluate SMILES against state-of-the-art supervised multi-instance learning algorithms on five popular benchmark datasets, twenty text datasets from the 20Newsgroups corpus and three datasets for the task of biocreative text categorization (see the Appendix for detail).

Since there is no unsupervised MIL algorithm at present, to evaluate the proposed SMILES, we use three categories of

Algorithm	Average	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-SVM	77.9±N/A	87.4±N/A	83.6±N/A	58.2±N/A	78.4±N/A	82.2±N/A
MI-SVM	77.7±N/A	77.9±N/A	84.3±N/A	57.8±N/A	84.2±N/A	84.3±N/A
MI-Kernel	81.2±2.0	88.0±3.1	89.3±1.5	60.3±2.8	84.2±1.0	84.3±1.6
EM-DD	76.5±4.4	84.9±4.4	86.9±4.8	60.9±4.5	73.0±4.3	77.1±4.3
mi-Graph	82.8±3.7	88.9±3.3	90.3±3.9	62.0±4.4	86.0±3.7	86.9±3.5
MI-VLAD	80.4±4.0	87.1±4.3	87.2±4.2	62.0±4.4	81.1±3.9	85.0±3.6
mi-FV	81.5±4.1	90.9±4.2	88.4±4.2	62.1±4.9	81.3±3.7	85.2±3.6
MI-SDB	87.4±3.7	**93.1±4.0**	91.2±4.1	78.9±3.4	86.5±4.2	87.5±3.0
BDR	84.5±3.5	92.4±2.7	90.3±5.2	62.8±3.4	86.9±3.6	90.2±2.9
mi-Net	80.8±3.8	88.9±3.9	85.8±4.9	61.3±3.5	82.4±3.4	85.8±3.7
MI-Net	81.2±3.8	88.7±4.1	85.9±4.6	62.2±3.8	83.0±3.2	86.2±3.4
MI-Net (DS)	82.3±3.8	89.4±4.2	87.4±4.3	63.0±3.7	84.5±3.9	87.2±3.2
MI-Net (RC)	81.6±4.2	89.8±4.3	87.3±4.4	61.9±4.7	83.6±3.7	85.7±4.0
Attention	81.4±3.5	89.2±4.0	85.8±4.8	61.5±4.3	83.9±2.2	86.8±2.2
Gated-Attention	81.3±3.3	90.0±5.0	86.3±4.2	60.3±2.9	84.5±1.8	85.7±2.7
B-Graph	82.2±3.0	89.7±3.7	87.1±2.8	64.0±4.1	82.9±2.2	87.5±2.4
SMILES	**92.9±2.1**	92.7±1.2	**96.2±1.6**	**85.5±4.3**	**92.0±1.5**	**98.2±2.0**

Table 1: Mean and standard error (when available) of classification accuracy (in %) for benchmark MIL datasets. The best results in each column are shown in bold. Higher accuracies are better.

supervised baselines: (i) the instance space approaches include mi-SVM and MI-SVM (Andrews, Tsochantaridis, and Hofmann 2002), EM-DD (Zhang and Goldman 2001), MI-VLAD and mi-FV (Wei, Wu, and Zhou 2017); (ii) the bag space methods include MI-Kernel (Gärtner et al. 2002), mi-Graph (Zhou, Sun, and Li 2009), BDR (Huang et al. 2022) and MI-SDB (Feng et al. 2021); (iii) we also compare with the embedding space methods mi-Net and MI-Net (Wang et al. 2018), Attention Neural Network and Gated Attention Neural Network (Ilse, Tomczak, and Welling 2018), and B-Graph (Pal et al. 2022), these methods use neural networks or attention to learn embeddings of the bags.

For the baselines, we set the hyper-parameters as suggested by their authors. For SMILES, we report the mean 10-fold cross validation accuracy after 5 runs followed by a linear SVM. The linear SVM is trained by applying cross validation on training data folds and the best mean accuracy is reported. We conduct experiment with the values of the number of message passing layers, the number of epochs, batch size, the parameter C of SVM, the threshold ϵ, augmentation ratio c and temperature τ in the sets $\{2, 4, 8, 12\}$, $\{10, 20, 40, 100\}$, $\{32, 64, 128, 256\}$, $\{10^{-3}, \ldots, 10^{2}, 10^{3}\}$, $\{0.1, \ldots, 0.5\}$, $\{10\%, \ldots, 50\%\}$ and $\{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$ respectively. The hidden dimension of layer is set to 128. Based on the Remark, the augmentation strategies include dropping, masking, replacing, and randomizing instances.

MIL Benchmark Datasets

We first evaluate our proposed framework on the benchmark datasets MUSK1, MUSK2 (Dietterich, Lathrop, and Lozano-Pérez 1997) for drug activity prediction, and FOX, TIGER, and ELEPHANT (Andrews, Tsochantaridis, and Hofmann 2002) for image classification. Table 1 shows the MIL result of each algorithm. It is observed that all the deep learning approaches are not well suited for these datasets as they are composed of precomputed features and the size of the bags are relatively small. But surprisingly, SMILES not only outperforms all deep supervised models, but achieves the state-of-the-art results with respect to the traditional supervised MIL algorithms on these small datasets. For example, the accuracy of SMILES is **25.5%** higher than B-Graph, the best deep baseline, and **6.6%** higher than MI-SDB, the best traditional algorithm, in data FOX. Tabel 1 lists the average accuracy of 5 benchmark datasets, from which SMILES achieves the best performance as well.

20 NewsGroups

In this section, we conduct the experiment on corpus data 20 NewsGroups (Zhou, Sun, and Li 2009). It contains posts from newsgroups on 20 subjects. When one of the subjects is selected as the positive class, all 19 other subjects are used as the negative class. The bags are collections of posts from different subjects. The classification accuracy with comparison to supervised models is summarized in Table 2.

From Table 2 we observe that all neural network based models outperform the classical MIL models on average in this task. This result suggests that using neural network can get better performance on these corpus data. In Table 2, we can find that our method significantly outperforms all the baselines for all the cases except for the *misc.forsale*, *sci.electronics* and *rec.motorcycles*. For example, SMILES outperforms the second best algorithm **13.4%** on *talk.politics.mideast*, **11.4%** on *sci.crypt*, **11.2%** on *rec.autos*, **10.0%** on *sci.space*, **10.4%** on talk.politics.guns, and **10.8%** on *talk.politics.misc*. And the average classification accuracy of all 20 multi-instance datasets indicate that our method outperforms others baselines, including MI-Kernel, mi-Graph, miFV, MI-SDB, mi-Net, MI-Net and its variants, and B-graph, with about **12.5%** improvement in

Algorithm	SMILES	MI-Kernel	mi-Graph	mi-FV	MI-SDB	mi-Net	MI-Net	MI-Net (DS)	MI-Net (RC)	B-Graph
alt.atheism	**89.1±2.4**	60.2±3.9	65.5±4.0	84.5±1.4	85.7±1.2	83.1±2.3	84.7±1.8	84.4±2.0	83.6±1.5	88.5±2.2
comp.graphics	**90.3±1.3**	47.0±3.3	77.8±1.6	59.6±5.8	74.6±1.1	81.7±0.6	82.0±1.5	81.9±0.5	81.5±0.9	80.1±3.2
comp.os.ms-windows.misc	**72.5±3.7**	51.0±5.2	63.1±1.5	61.3±1.2	67.1±3.2	70.4±1.7	70.7±1.1	70.9±1.1	70.7±1.4	71.9±3.6
comp.sys.ibm.pc.hardware	**80.0±2.7**	46.9±3.6	59.5±2.7	65.9±3.4	67.8±2.8	79.0±1.8	78.6±1.0	78.3±1.3	78.5±1.0	75.5±3.4
comp.sys.mac.hardware	**85.0±1.1**	44.5±3.2	61.7±4.8	65.9±2.4	65.7±2.3	79.4±1.6	79.1±1.5	79.7±1.1	79.2±1.9	79.2±3.1
comp.windows.x	**91.2±2.4**	50.8±4.3	69.8±2.1	76.9±3.5	79.0±1.9	79.9±1.8	80.9±1.9	80.1±1.1	81.2±2.7	86.1±2.7
misc.forsale	68.4±3.6	51.8±2.5	55.2±2.7	56.6±2.7	57.2±2.4	67.1±0.9	66.7±1.2	66.0±1.6	67.2±1.2	**75.8±3.5**
rec.autos	**90.1±2.7**	52.9±3.3	72.0±3.7	66.7±5.4	77.5±2.3	76.5±1.2	76.9±1.6	76.4±1.6	76.1±1.6	78.9±3.3
rec.motorcycles	72.5±2.7	50.6±3.5	64.0±2.8	80.0±1.6	**85.8±1.9**	83.4±1.1	84.2±1.0	83.5±1.5	83.3±1.3	85.5±2.4
rec.sport.baseball	**95.0±3.6**	51.7±2.8	64.7±3.1	78.0±2.7	82.1±2.5	86.0±1.6	86.7±1.7	85.7±2.5	87.1±1.4	83.5±3.1
rec.sport.hockey	**93.4±2.6**	51.3±3.4	85.0±2.5	82.4±4.2	90.8±1.8	89.0±1.7	90.2±1.4	91.1±1.6	89.8±1.1	90.0±2.3
sci.crypt	**93.3±3.1**	56.3±3.6	69.6±2.1	76.1±3.0	78.6±2.1	79.5±1.4	77.9±1.5	77.8±2.6	78.6±2.3	81.9±3.7
sci.electronics	87.5±3.1	50.6±2.0	87.1±1.7	55.5±1.4	90.1±2.2	92.1±0.8	**93.2±0.4**	92.7±0.5	93.1±0.7	91.4±2.9
sci.med	**90.5±2.4**	50.6±1.9	62.1±3.9	78.4±1.8	78.2±2.7	85.5±0.9	84.2±0.7	84.7±1.3	83.8±1.4	79.8±3.2
sci.space	**98.9±2.9**	54.7±2.5	75.7±3.4	81.7±1.5	83.9±1.5	79.8±1.3	79.5±2.8	80.1±2.6	80.3±2.6	88.9±2.6
soc.religion.christian	**87.5±3.4**	49.2±3.4	59.0±4.7	81.5±2.2	81.4±2.0	79.9±1.5	80.7±1.7	80.1±1.4	80.5±2.0	79.6±3.6
talk.politics.guns	**88.6±2.8**	47.7±3.8	58.5±6.0	74.7±1.9	75.8±3.5	76.1±1.9	78.2±1.8	77.0±2.4	77.3±1.0	77.7±4.9
talk.politics.mideast	**97.4±3.9**	55.9±2.8	73.6±2.6	79.9±3.4	80.6±2.0	83.9±1.0	84.0±1.2	83.8±1.0	83.3±2.0	82.0±3.4
talk.politics.misc	**88.8±4.5**	51.5±3.7	70.4±3.6	69.9±2.1	72.3±2.8	76.5±1.5	75.8±2.3	76.8±2.2	75.6±1.9	78.0±5.0
talk.religion.misc	**82.5±2.2**	55.4±4.3	63.3±3.5	74.0±3.8	73.9±2.6	74.4±1.5	76.2±1.7	76.2±1.5	74.3±1.2	80.1±3.6
Average	**87.2±2.8**	51.5±3.3	67.9±3.1	72.5±2.7	77.4±2.2	80.1±1.4	80.5±1.5	80.3±1.5	80.2±1.5	81.7±3.2

Table 2: Mean and std. error of classification accuracy (in %) along with average accuracy of the algorithms for the 20 Newsgroups datasets. The best results in each row are shown in bold. Higher accuracies are better.

Augmentation	MUSK2	FOX	TIGER	ELEPHANT
NoAug	88.3±1.7	80.5±1.3	90.1±2.9	92.2±2.3
20% Drop	89.9±3.2	83.2±2.8	90.3±3.3	94.5±3.1
20% Mask	**95.1±2.0**	**85.3±3.2**	**91.1±2.4**	95.0±2.4
20% Replace	92.7±3.1	**85.3±1.6**	91.0±1.5	**96.8±2.0**
20% Random	91.6±3.8	81.4±2.2	90.7±3.0	93.9±1.5

Table 3: Ablation study of bag augmentations on 4 datasets.

Structure	MUSK2	FOX	TIGER	ELEPHANT
NoGraph	80.2±1.4	65.8±2.1	62.1±4.2	69.2±3.6
WithGraph	**95.5±2.1**	**85.1±1.3**	**91.2±1.2**	**97.5±1.5**

Table 4: Ablation study of bag structure generation.

performance. For example, the average accuracy of SMILES is better than B-Graph by **5.5%**, MI-SDB by **9.8%**, mi-Graph by **19.3%**, respectively.

Ablation Analysis

Here we want to prove that the augmentation selection policy and the intensity of augmentations really matter to the final results. Note that we fixed one of the augmentation as NoAug (i.e. $\check{X} = X$) and all the other augmentation methods require a hyper-parameter "aug ratio" that controls the portion of instances/elements that are selected for augmentation. The "aug ratio" is set to a constant in every experiment (e.g., 20% by default). We perform an ablation study of different augmentation polices on four datasets MUSK2, FOX, TIGER, and ELEPHANT and intensities on FOX as shown in Table 3, Table 5 respectively. We conclude that: augmenting the bag data indeed boosts the performance of the proposed algorithm; the choice of "aug ratio" has a considerable effect on the final performance and the classification performance degenerates as the intensity of augmentation grows overly high. It is inappropriate to apply the same "aug ratio" to different augmentations.

Aug Ratio	Drop	Mask	Replace	Random
10%	83.0±1.3	84.5±2.1	81.3±3.2	**82.1±1.7**
20%	**83.2±2.8**	**85.3±3.2**	85.3±1.6	81.4±2.2
30%	82.5±2.0	85.0±4.1	**85.5±1.6**	82.0±0.3
40%	76.1±1.6	83.5±1.9	84.1±3.1	81.4±2.7
50%	73.7±1.4	81.3±1.6	82.8±2.0	79.5±1.8

Table 5: Ablation study of the *aug ratio* of bag augmentations on FOX dataset.

In addition, we want to analyze the impact of using structure on learning high-quality bag representation in unsupervised MIL. We conduct an additional ablation study over SMILES with (named WithGraph) or without (named NoGraph) bag structure awareness on four datasets MUSK2, FOX, TIGER, and ELEPHANT, shown in Table 4. From the Table 4, we observe that the accuracy of WithGraph is better than NoGraph **15.3%**, **19.3%**, **29.1%**, and **28.3%** on these datasets respectively, which demonstrates the importance of using structure recognition in multi-instance learning.

Robustness with Injected Noise

We compare SMILES to other sota supervised MIL methods with clean and injected noises in three datasets COMPONENT, PROCESS, and FUNCTION for the task of biocreative text categorization (Feng et al. 2021). To produce a noisy setting close to the real world, for each bag in dataset, there is η chance for it to be corrupted by noise. Specifically, all the instances x in the noise bag X is corrupted by $g(x; \delta)$, where δ is drawn from Gaussian distributions. Table 6 reports the classification accuracy of each method on

Datasets	Methods	Clean ($\eta = 0.0$)	Noise Rate (η) 0.1	0.2	0.3	0.4
COMPONENT	MI-Kernel	86.4 ± 1.1	86.1 ± 1.6	84.7 ± 2.6	83.4 ± 4.2	81.3 ± 2.4
	mi-Graph	90.4 ± 2.6	89.7 ± 2.0	88.3 ± 2.4	89.0 ± 1.1	88.5 ± 2.6
	mi-FV	91.0 ± 1.5	$\mathbf{90.8 \pm 2.1}$	89.1 ± 1.8	89.2 ± 2.7	88.9 ± 2.8
	mi-Net	89.5 ± 2.4	89.0 ± 3.0	88.4 ± 2.3	87.6 ± 1.7	87.3 ± 3.4
	MI-Net	89.8 ± 2.9	89.3 ± 2.7	88.8 ± 1.0	87.9 ± 2.5	87.3 ± 1.6
	Attention	90.8 ± 1.4	90.1 ± 3.2	89.3 ± 2.9	89.1 ± 2.1	88.3 ± 1.9
	Gated-Attention	$\mathbf{91.1 \pm 3.0}$	90.0 ± 1.6	89.3 ± 4.3	89.4 ± 2.4	88.5 ± 1.8
	B-Graph	87.4 ± 2.9	87.1 ± 1.4	86.4 ± 2.1	86.3 ± 1.0	86.0 ± 3.5
	MI-SDB	85.3 ± 4.0	84.6 ± 2.8	84.0 ± 3.6	82.6 ± 1.8	81.7 ± 2.9
	SMILES	$\mathbf{91.1 \pm 1.5}$	90.4 ± 2.4	$\mathbf{89.4 \pm 2.1}$	$\mathbf{91.0 \pm 3.0}$	$\mathbf{89.1 \pm 4.1}$
FUNCTION	MI-Kernel	91.5 ± 2.7	90.6 ± 3.2	89.7 ± 2.6	89.5 ± 1.7	87.4 ± 1.5
	mi-Graph	92.3 ± 3.3	91.9 ± 1.7	91.6 ± 4.1	91.7 ± 2.6	90.9 ± 2.8
	mi-FV	94.0 ± 2.1	93.8 ± 1.6	93.4 ± 1.4	93.7 ± 2.8	92.6 ± 2.5
	MI-Net	91.6 ± 3.2	91.6 ± 1.6	91.3 ± 2.4	90.8 ± 1.9	89.9 ± 1.4
	mi-Net	91.7 ± 2.1	91.6 ± 2.2	91.4 ± 4.0	91.0 ± 1.7	90.3 ± 2.8
	Attention	94.3 ± 1.7	94.0 ± 1.2	93.6 ± 1.6	92.6 ± 1.6	92.3 ± 1.2
	Gated-Attention	94.6 ± 4.0	$\mathbf{94.1 \pm 0.7}$	93.5 ± 1.2	92.9 ± 2.9	92.6 ± 2.5
	B-Graph	91.9 ± 2.3	91.5 ± 2.6	91.4 ± 3.1	91.4 ± 2.9	91.0 ± 2.2
	MI-SDB	81.7 ± 2.7	74.0 ± 2.3	72.1 ± 1.9	73.0 ± 3.5	71.1 ± 2.8
	SMILES	$\mathbf{94.7 \pm 2.7}$	$\mathbf{94.1 \pm 2.1}$	$\mathbf{93.7 \pm 2.3}$	$\mathbf{94.6 \pm 3.2}$	$\mathbf{92.7 \pm 1.2}$
PROCESS	MI-Kernel	92.3 ± 2.1	92.1 ± 2.3	91.5 ± 1.6	91.9 ± 2.8	90.8 ± 1.5
	mi-Graph	93.5 ± 2.7	93.1 ± 2.0	92.8 ± 1.4	91.0 ± 3.1	90.6 ± 2.1
	mi-FV	94.3 ± 1.7	94.0 ± 2.3	92.9 ± 3.0	92.6 ± 2.2	91.7 ± 2.6
	MI-Net	94.1 ± 1.5	93.8 ± 1.6	93.7 ± 2.5	93.5 ± 2.3	92.7 ± 1.9
	mi-Net	94.7 ± 3.0	94.2 ± 2.9	93.5 ± 1.6	93.0 ± 2.0	92.8 ± 2.4
	Attention	95.7 ± 2.7	95.5 ± 1.3	95.3 ± 2.2	94.9 ± 1.5	94.5 ± 2.3
	Gated-Attention	95.9 ± 1.7	95.6 ± 3.4	$\mathbf{95.5 \pm 1.3}$	95.3 ± 2.8	94.2 ± 2.0
	B-Graph	93.9 ± 2.9	93.5 ± 2.4	93.2 ± 3.2	92.4 ± 1.4	91.9 ± 2.7
	MI-SDB	84.8 ± 2.0	84.0 ± 1.7	83.1 ± 1.9	82.5 ± 3.5	75.5 ± 2.6
	SMILES	$\mathbf{96.2 \pm 3.6}$	$\mathbf{96.2 \pm 1.5}$	$\mathbf{95.5 \pm 2.1}$	$\mathbf{95.9 \pm 1.8}$	$\mathbf{95.8 \pm 2.2}$

Table 6: Test accuracies (%) of different methods on benchmark datasets with clean or noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The results (mean$\pm$std) are reported and the best results are boldfaced.

these three datasets. We can also observe that our proposed unsupervised bag representation method is clearly superior to other compared supervised MIL baselines.

In most cases, we produce a substantial improvement. On COMPONENT and FUNCTION, light noise (e.g., 10%, 20% noise) does not lead to much drop in the classification results. They are even comparable to other sota methods on the clean datasets. On PROCESS, when there is 30% noise, we improve the accuracy by the best sota baseline by nearly 0.6%, and when there is 40% noise, we improve the accuracy by the best sota method by 1.3%. This indicates that when the noise becomes more complex, the performance gap between the best supervised algorithm and SMILES further increase. Therefore, contrastive-based bag augmentation is a simple yet very effective trick to prevent noise. To summarize, compared with supervised baselines, SMILES produce improvements on the clean datasets. Moreover, when there are noises in data, SMILES is more robust than these supervised counterparts.

Conclusion

Self-supervised learning has seen success in various domains, but little progress has been made for the multi-instance data. In this paper, we proposed a self-supervised multi-instance learning method SMILES focused on learning the representations of bags that are effective in downstream classification tasks. SMILES provides a unified approach to meet a number of fundamental postulates including permutation invariance, structure-awareness and robustness, while conducts theoretical analysis towards understanding of our framework. Specifically, we augment the bags and train the encoder to maximize the agreement of two jointly sampled positive bag pairs to yield robust MIL model without label. To capture topological structures of bags, our framework learns graphs for the bags and these graphs are optimized together with message passing layers and ordered weighted averaging operator towards contrastive loss. Experiment results verify the state-of-the-art performance of our proposed framework in both generalizability and robustness.

Acknowledgments

This research would like to acknowledge the support from National Natural Science Foundation of China (62032013 and 62076143), Singapore Institute of Technology Ignition Grant (R-IE2-A405-0001), Innovative Talents of Higher Education in Liaoning Province (No. LR2020076) and Basic Research Operating Funds for National Defence Major Incubation Projects (No. N2116017).

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2002. Support vector machines for multiple-instance learning. *Advances in Neural Information Processing Systems*, 15.

Chevaleyre, Y.; and Zucker, J.-D. 2000. Noise-tolerant rule induction from multi-instance data. In *Proceedings of the ICML-2000 workshop on Attribute-Value and Relational Learning: Crossing the Boundaries*.

Chu, Y.; Yue, X.; Yu, L.; Sergei, M.; and Wang, Z. 2020. Automatic image captioning based on ResNet50 and LSTM with soft attention. *Wireless Communications and Mobile Computing*, 2020: 1–7.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1): 31–71.

Feng, L.; Shu, S.; Cao, Y.; Tao, L.; Wei, H.; Xiang, T.; An, B.; and Niu, G. 2021. Multiple-instance learning from similar and dissimilar bags. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 374–382.

Gärtner, T.; Flach, P. A.; Kowalczyk, A.; and Smola, A. J. 2002. Multi-instance kernels. In *International Conference on Machine Learning*, 179–186.

Huang, S.; Liu, Z.; Jin, W.; and Mu, Y. 2022. Bag dissimilarity regularized multi-instance learning. *Pattern Recognition*, 126: 108583.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*.

Lin, T.; Xu, H.; Yang, C.; and Xu, Y. 2022. Interventional multi-instance learning with deconfounded instance-level prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2): 1601–1609.

Luengo, J.; Sánchez-Tarragó, D.; Prati, R. C.; and Herrera, F. 2021. Multiple instance classification: Bag noise filtering for negative instance noise cleaning. *Information Sciences*, 579: 388–400.

Maron, H.; Fetaya, E.; Segol, N.; and Lipman, Y. 2019. On the universality of invariant networks. In *International Conference on Machine Learning*, 4363–4371. PMLR.

Pal, S.; Valkanas, A.; Regol, F.; and Coates, M. 2022. Bag graph: Multiple instance learning using bayesian graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Reed, C. J.; Yue, X.; Nrusimha, A.; Ebrahimi, S.; Vijaykumar, V.; Mao, R.; Li, B.; Zhang, S.; Guillory, D.; Metzger, S.; et al. 2022. Self-supervised pretraining improves self-supervised pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2584–2594.

Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; Wei, Y.; Huang, W.; and Huang, J. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33: 12559–12571.

Von Kügelgen, J.; Sharma, Y.; Gresele, L.; Brendel, W.; Schölkopf, B.; Besserve, M.; and Locatello, F. 2021. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34: 16451–16467.

Wagstaff, E.; Fuchs, F.; Engelcke, M.; Posner, I.; and Osborne, M. A. 2019. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, 6487–6494. PMLR.

Wagstaff, E.; Fuchs, F. B.; Engelcke, M.; Osborne, M. A.; and Posner, I. 2022. Universal approximation of functions on sets. *JMLR*, 23(151): 1–56.

Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74: 15–24.

Wei, X.; Wu, J.; and Zhou, Z. 2017. Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4): 975–987.

Yager, R. R. 1988. On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1): 183–190.

Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S.; Ji, X.; and Ye, Q. 2021. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5330–5339.

Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; and Smola, A. J. 2017. Deep sets. *Advances in Neural Information Processing Systems*, 30.

Zhang, L.; Tozzo, V.; Higgins, J.; and Ranganath, R. 2022. Set norm and equivariant skip connections: Putting the deep in deep sets. In *International Conference on Machine Learning*, 26559–26574. PMLR.

Zhang, Q.; and Goldman, S. A. 2001. EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems*, 14.

Zhang, W. 2021. Non-i.i.d. multi-instance learning for predicting instance and bag labels with variational auto-encoder. In Zhou, Z.-H., ed., *International Joint Conference on Artificial Intelligence*, 3377–3383. International Joint Conferences on Artificial Intelligence Organization.

Zhang, Z.; Liu, Q.; Wang, H.; Lu, C.; and Lee, C.-K. 2021. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34: 15870–15882.

Zhao, Y.; Wang, Y.; Wang, Z.; and Zhang, C. 2021. Multi-graph Multi-label Learning with Dual-granularity Labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2327–2337.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-i.i.d. samples. In *International Conference on Machine Learning*.