Learnable Bag Similarity Based Deep Multi-Instance Network for Breast Cancer Diagnosis

Rui Cheng[1,2,3], Liming Yuan[1,2,3*], Haixia Xu[1,2,3], Zhenliang Li[1,2,3] and Xianbin Wen[1,2,3]
[1]School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
[2]Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin 300384, China
[3]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin 300384, China
Email: ruicheeng@outlook.com; *yuanliming@tjut.edu.cn;
xuhaixia_xhx@163.com; ustarlee@outlook.com; xbwen@tjut.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-7281-6215-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/BIBM49941.2020.9313600

***Abstract*—Computer-aided diagnosis for breast cancer is an important and challenging research problem in medical image analysis. The main difficulty lies in that only image-level labels rather than fine-grained patch-level labels can be gotten for medical images in general. This situation fits well with the settings of Multi-Instance Learning (MIL). Following this line of research, Bag Similarity Network (BSN) uses the inter-bag similarities to learn the relationship between bags (images) and achieves a good performance in automatic diagnosis of breast cancer. Nevertheless, the inter-bag similarities are pre-defined rather than learnable. In this paper, we propose a Learnable Bag Similarity Network for deep MIL, called LBSN, to aid breast cancer diagnosis. To implement automatic similarity learning, LBSN first extracts fixed numbers of global representations of each bag using the attention mechanism, and then employs channel and spatial attention on similarity matrices for learning the relation of bags and that of instances, respectively. The experimental results on a publicly available breast cancer data set demonstrate that the proposed LBSN outperforms BSN by a large margin in terms of classification accuracy.**

***Index Terms*—Multi-instance learning, Deep learning, Bag similarity, Channel attention, Spatial attention**

I. INTRODUCTION

In histopathology, the cancer refers to malignant tumor originating from epithelial tissue and it is the most common type of malignant tumor. This disease progresses quickly with high mortality. It has caused great harm to the patient's body, psychology and economy, and has become a public health issue deserving people's attention [1]. So timely intervention through large-scale screening in early detection is important, however, it takes a lot of manpower to annotate the patients' H&E images finely, and it is difficult to ensure that the experienced doctors don't make mistakes in the labeling process.

Due to the difficulty in labeling, fully supervised learning is often not applicable for medical image analysis. Alternatively, some weakly supervised learning algorithms have been proposed to solve various medical image problems. Different from supervised learning algorithms, weakly supervised learning algorithms can learn from coarse-labeled images.

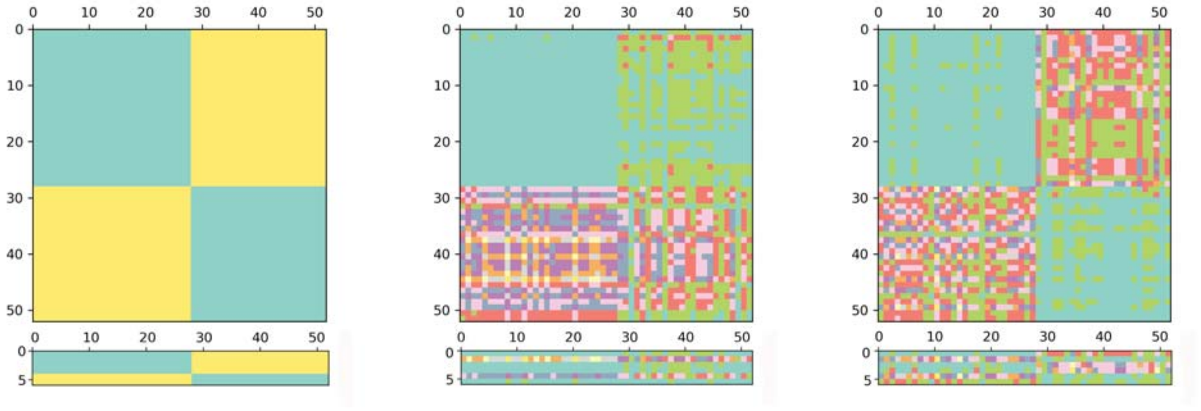
Multi-Instance Learning (MIL) [2] is a typical weakly supervised learning. In MIL, the training dataset is composed of bags, where each bag contains multiple feature vectors (called instances in the MIL terminology). In this learning framework, each bag has its own label, but the labels of individual instances are unknown [3]. Generally, MIL regards a medical image as a bag, with an image split into lots of patches regarded as instances. If there is at least a positive instance (cancerous patch), the bag will be labeled positive, otherwise it is negative [4–6].

MIL algorithms can be divided in to three groups: instance-space paradigm, bag-space paradigm and embedded-space paradigm based on the type of information extracted (instance-level or bag-level information) and how it is represented (implicitly or explicitly) [3, 7]. Instance-space paradigm learns an instance classifier firstly, then aggregates the instance-level results to perform the bag-level classification. Bag-space paradigm and embedded-space paradigm extract information from the whole bag. The difference between these two paradigms lies in the method of extracting bag-level information. Bag-space paradigm computes bag-to-bag distance/similarity implicitly, while embedded-space paradigm embeds a bag into another feature space explicitly [7].

Deep neural networks have been applied to MIL widely and effectively. Wang et al. [7] proposed two methods named mi-Net and MI-Net. They both optimize instance feature learning, bag feature learning, instance classification, and bag classification in a fully end-to-end manner via backpropagation. Ilse et al. [8] made improvements on the MI-Net pooling method, and applied the attention mechanism to the MIL pooling. This pooling method gives different weights to instances based on the contribution of different instances. However, these deep MIL methods only focus on instance-level information without considering bag-level information. The BSN method proposed by Wang et al. [9] is the first study that integrates bag similarity with deep neural network, and achieves good performance. However the Hausdroff pooling method in BSN is a pre-defined pooling method. If the target bag is a positive bag, the max-max pooling will extract inappropriate features from the matrix of the negative reference bag then cause a decrease

*Corresponding author

This work was supported by the National Natural Science Foundation of China (61472278), the Major Project of Tianjin (18ZXZNGX00150), the Natural Science Foundation of Tianjin (18JCYBJC84800), the Scientific Research Key Project of Tianjin Municipal Education Commission (2017ZD13), and the Scientific Research Project of Tianjin Municipal Education Commission (2017KJ255).



(a) Groundtruth

(b) BSN

(c) LBSN

Fig. 1. Comparison between the learned similarity matrices by BSN and LBSN on UCSB dataset. The top row and bottom row show the similarity matrices of training bags and test bags, respectively. LBSN has more significant diagonal-blockness of similarity matrix than BSN.

on performance.

In this study, we propose a learnable bag similarity network called LBSN and use the channel-spatial attention mechanism to extract important imformation which satisfies the needs of the classifier. A comparison between the bag similarity obtained by BSN and the proposed method is shown in Fig. 1, where it can be seen that the different block boundaries of bag similarity matrix are more obvious compared to BSN. Specifically, we obtain a new instance set which is extracted by multiple-perspective query vectors of attention mechanism from original bag and regard these sets as reference bags. Then channel attention and spatial attention are used to obtain the effect of different reference bags on the target bag and the relationship of different instances between the target bag and reference bags, respectively. Our study is the first work to integrate channel attention and spatial attention with the MIL algorithm, and has also achieves excellent performance on the histopathology data set.

The rest of this article is oraganized as follows. Section II briefly reviews previous studies on deep MIL. In Section III, we propose a learnable bag similarity network. The experimental results on a breast cancer data set set are presented in Section IV. Finally, in Section V, we conclude the article with some future studies.

II. RELATED WORK

A. Deep Multi-Instance Learning

Deep MIL networks have improved the classification performance significantly, compared to previous traditional shallow MIL networks. Wu et al.[10] observe the generic existence of the multiple instance assumption in object and keyword candidates, and incorporate deep learning into a weakly supervised learning framework in a principled manner. The MIL-based classifier that aggregats the instance-level predictions proposed by Hou et al. [11] can improve the performance of the classifier greatly. MINN can be regarded as the first general framework in the deep MIL neural network. The main contributions of MINN are proposing two extremely fast and scalable methods (mi-Net and MI-Net) for MIL and introducing deep supervision and residual connections for MIL[7].

B. Bag Similarity Network

Some methods that do not make explicit assumptions about the instances or the concepts, but only suppose that bags of same class are similar to each other and then learn from the distance or similarity between bags. Such methods include Citation-kNN [12], bag kernels [13], and bag dissimilarities [14, 15]. Cheplygina et al.[16] express each bag by its dissimilarities with training bags, thus the MIL problem is converted to a supervised learning problem that can use any classifier, and they point out that the dissimilarity based on the averag of the minimum instance distances between bags shows good performance in the most real-life datasets. BSN proposed by Wang et al.[9] is the first study that integrates the bag similarity with MIL neural networks, and adopts a decoupled training scheme. First, training MI-Net network to obtain all instance features; then updating the features of target instance according to the features of reference instances.

C. Attention

Attention model was originally used in machine translation [17], and has become an important concept in the field of neural networks. After that it has been used in natural language processing, statistical learning, speech and computer vision widely. The attention model on the CNN network proposed by Hu et al. [18] has achieved great improvement compared to the traditional CNN networks. They proposed a novel architectural unit, which is termed the ”Squeeze-and-Excitation” (SE) block that focuses on the channel-wise feature responses by modeling interdependencies between channels explicitly. The CBAM module proposed by Woo et al. [19] considers convolutional features in more details. For intermediate feature maps, the module generates attention maps along the channel and space in turn, and assigns the attention maps to the feature maps to achieve feature refinement. Based on the attention mechanism, Ilse et al. [8] proposed a method aimed at combining interpretability with the MIL method for greater flexibility. It also proves that the application of the Fundamental Theorem of Symmetric Functions provides a general process for modeling the bag label probability.

III. PROPOSED METHOD

A. Overview

The proposed method not only solves the problem from the new perspective of the bag, but also uses the importance of the relationship between bags to select the bag similarity instead of the Hausdorff pooling method. The channel attention module obtains the effect of different reference bags on the target bag. For the target bag, the bag with the same label in the reference bags will be more effective than the bag with other labels. In the spatial attention module, the key instance in the bag is given a higher degree of importance. Our network is divided into two steps for training. The first step is to learn the instance-level representations of reference bags by Attention network (AttNet)[8] architecture as shown in Fig. 2, the second step is to classify the bag-level information that is computed from the target bag and the reference bags as shown in Fig. 3.

B. MIL Formulation

We define a set of bags $X = \{X_1, X_2, ..., X_N\}$ and instance features of ith bag $X_i = \{x_{i1}, x_{i2}, ..., x_{im_i}\}$, $x_{ij} \in \mathbb{R}^{d\times 1}$, where N and m_i denote the number of bags and the number of instances in bag X_i, respectively. The label of bag X_i and instance x_{ij} are $Y_i \in \{0,1\}$ and $y_{ij} \in \{0,1\}$. Generally, the SMI assumption is denoted as follows:

$$Y_i = \begin{cases} 0, & \text{if } \sum_{j=1}^{mi} y_{ij} = 0 \\ 1, & \text{else.} \end{cases} \tag{1}$$

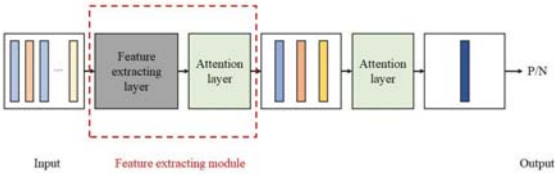


Fig. 2. The framework of AttNet. The AttNet is composed of a feature extracting module, a pooling module and a classifier. The purpose of the feature extracting module is to obtain a new instance representation (different differentiation of cells). The pooling layer (the attention layer) can aggregate instances embedded into another feature space.

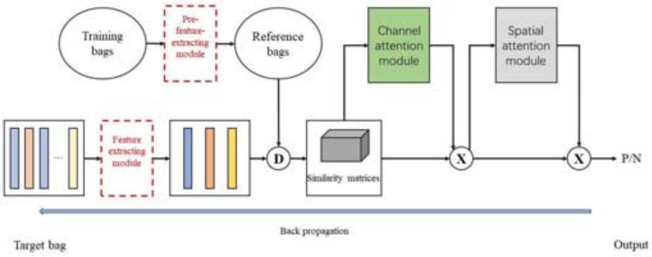


Fig. 3. The framework of LBSN. First, we obtain reference bags by pre-trained AttNet, and then the target bag is transformed into a new instance representation by a new feature extracting module. Finally, the similarity matrices calculated from the target bag and the reference bags will be sent to the bag-level classifier through the channel attention module and the spatial attention module.

C. AttNet

In order to obtain the new representation, we train a novel AttNet structure network which is different with the original AttNet[8]. The proposed AttNet is formalized as

$$P(X_i) = f(\text{AttPooling}_2(\text{AttPooling}_1(\text{FE}(X_i)))), \tag{2}$$

where $P(X_i) \in [0,1]$ is the probability of $Y_i = 1$, FE (feature extracting layer) is a transformation of instances and f is a classifier. AttPooling is an attention-based MIL pooling defined as

$$\text{AttPooling}(X_i) = X_i\text{softmax}(\tanh(X_i^T V)W), \tag{3}$$

which is the weighted average of the input set X_i, where V and W are the learnable parameters of layer, the difference between AttPooling_1 and AttPooling_2 is the size of W. FE is parameterized by a permutation equivariant neural network (e.g., generally use fully-connected neural network or CNN, specific to the type of data), and either of AttPooling and f by a fully-connected layer. It is worth noting that when $X_i = \{x_{i1}, x_{i2}, ..., x_{im_i}\}$ is transformed into $X_i = \{x_{i1'}, x_{i2'}, x_{i3'}\}$ by a feature extracting module, $x_{ij'}$ devotes the new weighted averaged instance.

D. LBSN

LBSN is composed of a pre-defined AttNet, an attention module and a classifier. To compute the similarity matrices between target bag X_t and reference bags, we construct the LBSN which is learnable from different reference bags. The target bag $X_t = \{x_{t1}, x_{t2}, ..., x_{tm_t}\}$ is transformed into $X_t = \{x_{t1'}, x_{t2'}, x_{t3'}\}$ by a new feature extracting module. Then, we calculate the inner product $\text{s}(x,y) = x^T y$ of the instances of X_t and each reference bag to form the bag similarity matrix:

$$\begin{aligned} S &= [\mathrm{s}(X_t, X_{R_1}), ..., \mathrm{s}(X_t, X_{R_N})], \\ S' &= \mathrm{M}_c(S) \bigotimes S, \\ S^{''} &= \mathrm{M}_s(S') \bigotimes S', \\ S^{'''} &= \sum_{i=1,j=1}^{3} S^{''}_{ij} \\ P(X_t) &= f'(S'''), \end{aligned} \tag{4}$$

where $S \in \mathbb{R}^{N\times3\times3}$ denotes the bag similarity matrices, in the meantime we regard the first dimension of bag similarity matrices as channel, $\bigotimes$ denotes the element-wise multiplication, M_c denotes the channel attention operation and M_s denotes the spatial attention operation, f' is a bag-level classifier. Lastly, $P(X_t) \in [0, 1]$ predicts the probability of $Y_t = 1$.

Channel attention module. We produce a channel attention map by exploiting the bag relationship between a target bag and reference bags. The channel attention module is shown in Fig. 4. Since each channel of the feature map is considered as a feature detector, channel attention focuses on which reference bags are meaningful bags. In order to calculate the channel attention effectively, average-pooling has been adopted to represent each channel information. Woo et al. [19] not only adopt the average-pooling, but also argue that max-pooling gathers another clue about distinctive object features to infer channel-wise attention. In MIL, if we adopt max-pooling as one of aggregating channel information, it will loss information from other channel information. Thus, we only use average-pooling as the pooling method and describe the detailed operation below.

We first aggregate channel information of a similarity matrix by using average-pooling generating a scriptor: S^c_{avg}, which denotes average-pooling feature. Then average-pooling feature is sent to a shared conv to produce $\mathrm{M}_c(S) \in \mathbb{R}^{N\times1\times1}$. The shared conv is composed of two convolutional layers. To reduce parameters overhead, we set output channels of first convolutional layer as half of input channels and then we recover the second convolutional layer's output channel to the number of reference bags.

$$\begin{aligned} \mathrm{M}_c(S) &= \sigma(\mathrm{Conv}(\mathrm{AvgPool}((S))) \\ &= \sigma(W_1(W_0(S^c_{avg}))), \end{aligned} \tag{5}$$

where σ denotes sigmoid function, $W_0 \in \mathbb{R}^{N\times N/2}$, and $W_1 \in \mathbb{R}^{N/2\times N}$. The two Convolutional layers both share weights W_0 and W_1, and the ReLU activation function is followed by W_0.

Spatial attention module. We generate spatial attention by using the spatial relationship between the new aggregating instances. Unlike channel attention, the spatial attention focuses on which is the important part in similarity matrix and it is a supplement to channel information. The spatial attention module is shown in Fig. 5. To compute spatial attention, we apply average-pooling operation along the channel axis firstly, and then highlight the important area. Lastly, we generate a spatial attention map $\mathrm{M}_s(S) \in \mathbb{R}^{3\times3}$. Specifically, we aggregate channel information of a channel-refined similarity matrix by using average-pooling generating a scriptor: $S^s_{avg} \in \mathbb{R}^{1\times3\times3}$.

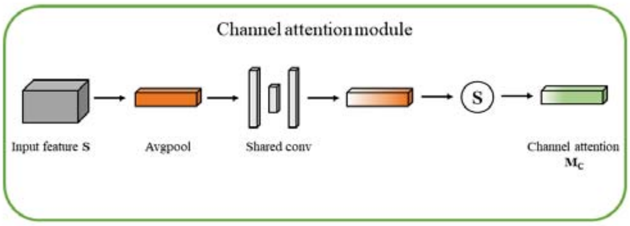


Fig. 4. Diagram of the channel attention module. As illustrated, the channel attention module uses the average-pooling output through the shared convolutional network.

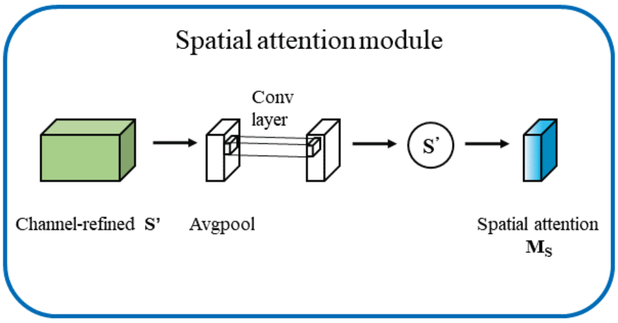


Fig. 5. Diagram of the spatial attention module. As illustrated, the spatial attention module uses the average-pooling output along the channel axis through a standard convolutional layer.

$$\begin{aligned} \mathrm{M}_s(S) &= \sigma(f^{1\times1}([\mathrm{AvgPool}(S)])) \\ &= \sigma(f^{1\times1}([S^s_{avg}])), \end{aligned} \tag{6}$$

where σ denotes the softmax function and $f^{1\times1}$ represents a convolutional operation with the size of 1×1.

IV. Experiments

A. Dataset

We evaluate our method on hematoxylin and eosin (H&E) stained TMA image data set.

Breast cancer data set[20]: This public data set consists of 58 TMA image excerpts of 896×768 pixel size taken from breast cancer patients, 32 of which are at benign status, and 26 are malignant. If an image contains breast cancer cells, it will be labled malignant, otherwise it is benign. The learning task is to classify images as benign and malignant . We split each image into an equal-sized 7×7 patch. In this data set, we use the model proposed in [7] for the feature extracting.

B. Experimental setup

In our experiments, we use default hyper-parameters for training LBSN with AttNet, including learning rate(LR), weight decay(WD), and the momentum(M) are given in

TABLE I
HYPER-PARAMETERS SETTING ON DATA SET

Dataset	**LR**	**WD**	**M**
Ucsb breast	0.0005	0.005	0.9

TABLE I. For UCSB breast data set, the training process is terminated at 100 epochs, the network is optimized by stochastic gradient descent techniques. All results are obtained by 10-fold cross-validation.

C. Ablation study

To evaluate the effectiveness of different components in the proposed method, we conduct ablation studies. We try different combinations of the following: (1) New AttNet (2) New AttNet + BSN (3) New AttNet + LBSN (channel attention module adopts two pooling method: avg+max) (4) our method.

The results of all these configurations are given in TABLE II. Comparing (1) to (2), (1) to (3) and (1) to (4), it is noted that the use of the bag-level feature improves accuracy. Similarly, when comparing (2) to (4), the result of accuracy in 0.8% gain. The comparison between (3) and (4) shows the average-pooling method is better than two-pooling method.

D. Comparision with State-of-the-Art methods

TABLE III demonstates the comparison between our proposed method and other state-of-the-art methods including MI-Net [7], AttNet [8], Gated AttNet [8], BSN [9]. We reimplement all the previous methods based on the literatures and open source codes. From the TABLE III, we can observe that our approach is better than these methods.

MI-Net [7] is the first general framework in deep MIL neural network. The MI-Net method utilizes the embedded instance-level imformation as the feature and adopts the max-pooling method to aggregat instances into bag-level representation. However, this simple pooling method is not suitable for data set with a large amount of training data. Especially the max-pooling method lacks interpretability, and the actual meaning represented by the maximum value of each dimension in the instances cannot be explained from the MIL problem. This pooling method does not consider the contribution of different instances to the classifier, thereby reducing its performance. In histopathology data set, the maximum value of the different dimensions in the instance does not represent which instance is the diseased area.

TABLE II
THE RESULTS OF ABLATION STUDY.

Method	Accuracy
New AttNet	0.874 ± 0.129
New AttNet + BSN	0.879 ± 0.126
New AttNet + LBSN(avg+max)	0.878 ± 0.132
LBSN(ours)	**0.887** ± 0.126

TABLE III
DIFFERENT METRICS ON BREAST CANCER DATA SET.

Method	Accuracy
MI-Net	0.832 ± 0.151
AttNet	0.867 ± 0.127
Gated AttNet	0.874 ± 0.137
BSN	0.749 ± 0.181
LBSN(ours)	**0.887** ± 0.126

AttNet and Gated-AttNet [8] methods are flexible and interpretable MIL methods. This kind of trainable MIL pooling based on the attention mechanism takes into account the importance of different instances in the bag, and is parameterized through a neural network then assigned to different instances. It is important for the histopathology data set to assign higher weights to diseased area patches. However, in some applications, if there are dependencies among instances within a bag, this pooling method may inhibit the performance of the model.

The BSN method [9] considers the similarity of different bags in the training data to the target bag, but it performs poorly in the medical image histopathology datasets. Because the reference H&E images are highly similar, and the similarity of representations in BSN are not sufficiently discriminative. The second reason that the Hausdorff pooling method uses a predefined method for the selection of similarity, reference bags that are not important to the target bag may have a bad effect on the bag-level classifier.

Generally, LBSN outperforms all the other methods. Compared against the best results obtained by the existing methods, LBSN increases accuracy by 1.3% in UCSB breast data set. We conclude reasons for the outstanding performance in prediction accuracy as follows.

Firstly, the pre-trained AttNet extracts the new representation of bag which contains important instances. Secondly, the LBSN considers bag similarity representation with whole data set relations. Thirdly, the attention module is also a very effcient method to select bag similatries and relation of instances.

V. CONCLUSION

In this paper, we investigate a challenging clinical task of automatic prediction of cancer using histopathological images of breast cancer. To achieve this, we propose a deep neural network named LBSN based on MIL. This network is composed of the AttNet and the LBSN, and experimental results benefit from these. (1) AttNet selects instance-level features effectively, (2) LBSN produces more effective bag-level features. Our approach shows outstanding performance compared to the state-of-art methods. In the future, it would be meaningful to consider how to optimize our channel and spatial attention module and how to reduce time spent in the training process.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

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

[3] J. Amores, “Multiple instance classification: Review, taxonomy and comparative study,” *Artificial intelligence*, vol. 201, pp. 81–105, 2013.

[4] J. R. Foulds and E. Frank, “A review of multi-instance learning assumptions,” 2010.

[5] G. Liu, J. Wu, and Z.-H. Zhou, “Key instance detection in multi-instance learning,” 2012.

[6] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang, “Deep learning of feature representation with multiple instance learning for medical image analysis,” in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 1626–1630.

[7] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, “Revisiting multiple instance neural networks,” *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[8] M. Ilse, J. M. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” *arXiv preprint arXiv:1802.04712*, 2018.

[9] X. Wang, Y. Yan, P. Tang, W. Liu, and X. Guo, “Bag similarity network for deep multi-instance learning,” *Information Sciences*, vol. 504, pp. 578–588, 2019.

[10] J. Wu, Y. Yu, C. Huang, and K. Yu, “Deep multiple instance learning for image classification and auto-annotation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3460–3469.

[11] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, “Patch-based convolutional neural network for whole slide tissue image classification,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2016, pp. 2424–2433.

[12] J. Wang and J.-D. Zucker, “Solving multiple-instance problem: A lazy learning approach,” 2000.

[13] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, “Multi-instance kernels,” in *ICML*, vol. 2, no. 3, 2002, p. 7.

[14] D. M. Tax, M. Loog, R. P. Duin, V. Cheplygina, and W.-J. Lee, “Bag dissimilarities for multiple instance learning,” in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2011, pp. 222–234.

[15] M.-L. Zhang and Z.-H. Zhou, “Multi-instance clustering with applications to multi-instance prediction,” *Applied intelligence*, vol. 31, no. 1, pp. 47–68, 2009.

[16] V. Cheplygina, D. M. Tax, and M. Loog, “Multiple instance learning with bag dissimilarities,” *Pattern recognition*, vol. 48, no. 1, pp. 264–275, 2015.

[17] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[18] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[19] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[20] M. Kandemir, C. Zhang, and F. A. Hamprecht, “Empowering multiple instance histopathology cancer diagnosis by cell graphs,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 228–235.